NONPARAMETRIC LEAST SQUARES ESTIMATION OF A MULTIVARIATE CONVEX REGRESSION FUNCTION

Emilio Seijo and Bodhisattva Sen

Columbia University

July 10, 2010

Abstract

This paper deals with the consistency of the least squares estimator of a convex regression function when the predictor is multidimensional. We characterize and discuss the computation of such an estimator via the solution of certain quadratic and linear programs. Mild sufficient conditions for the consistency of this estimator and its subdifferentials in fixed and stochastic design regression settings are provided. We also consider a regression function which is known to be convex and componentwise nonincreasing and discuss the characterization, computation and consistency of its least squares estimator.

1 Introduction

Consider a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$, for $d \geq 1$, with nonempty interior and a regression model of the form

$$Y = \phi(X) + \epsilon \tag{1}$$

where X is a $\mathfrak{X}$-valued random vector, ϵ is a random variable with $\mathbf{E}(\epsilon|X) = 0$, and $\phi : \mathbb{R}^d \to \mathbb{R}$ is an unknown *convex* function. Given independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from such a model, we wish to estimate ϕ by the method of least squares, i.e., by finding

a convex function $\hat{\phi}_n$ which minimizes the discrete $\mathscr{L}_2$ norm

$$\left(\sum_{k=1}^{n} \left| Y_k - \psi(X_k) \right|^2 \right)^{\frac{1}{2}}$$

among all convex functions ψ defined on the convex hull of $X_1, \ldots, X_n$. In this paper we characterize the least squares estimator, provide means for its computation, study its finite sample properties and prove its consistency.

The problem just described is a nonparametric regression problem with known shape restriction (convexity). Such problems have a long history in the statistical literature with seminal papers like Brunk (1955), Grenander (1956) and Hildreth (1954) written more than 50 years ago, albeit in simpler settings. The former two papers deal with the estimation of monotone functions while the latter discusses least squares estimation of a concave function whose domain is a subset of the real line. Since then, many results on different nonparametric shape restricted regression problems have been published. For instance, Brunk (1970) and, more recently, Zhang (2002) have enriched the literature concerning isotonic regression. In the particular case of convex regression, Hanson and Pledger (1976) proved the consistency of the least squares estimator introduced in Hildreth (1954). Some years later, Mammen (1991) and Groeneboom et al. (2001) derived, respectively, the rate of convergence and asymptotic distribution of this estimator. Some alternative methods of estimation that combine shape restrictions with smoothness assumptions have also been proposed for the one-dimensional case; see, for example, Birke and Dette (2006) where a kernel-based estimator is defined and its asymptotic distribution derived.

Although the asymptotic theory of the one-dimensional convex regression problem is well understood, not much has been done in the multidimensional scenario. The absence of a natural order structure in $\mathbb{R}^d$, for $d > 1$, poses a natural impediment in such extensions. A convex function on the real line can be characterized as an absolutely continuous function with increasing first derivative (see, for instance, Folland (1999),

Exercise 42.b, page 109). This characterization plays a key role in the computation and asymptotic theory of the least squares estimator in the one-dimensional case. By contrast, analogous results for convex functions of several variables involve more complicated characterizations using either second-order conditions (as in Dudley (1977), Theorem 3.1, page 163) or cyclical monotonicity (as in Rockafellar (1970), Theorems 24.8 and 24.9, pages 238-239). Interesting differences between convex functions on $\mathbb{R}$ and convex functions on $\mathbb{R}^d$ are given in Johansen (1974) and Bronšteĭn (1978).

Recently there has been considerable interest in shape restricted function estimation in multidimension. In the density estimation context, Cule et al. (2010) deal with the computation of the nonparametric maximum likelihood estimator of a multidimensional log-concave density, while Cule and Samworth (2010), Schuhmacher et al. (2009) and Schuhmacher and Dümbgen (2010) discuss its consistency and related issues. Seregin and Wellner (2009) study the computation and consistency of the maximum likelihood estimator of convex-transformed densities. This paper focuses on estimating a regression function which is known to be convex. To the best of our knowledge this is the first attempt to systematically study the characterization, computation, and consistency of the least squares estimator of a convex regression function with multidimensional covariates in a *completely nonparametric* setting.

In the field of econometrics some work has been done on this multidimensional problem in less general contexts and with more stringent assumptions. Estimation of concave and/or componentwise nondecreasing functions has been treated, for instance, in Banker and Maindiratta (1992), Matzkin (1991), Matzkin (1993), Beresteanu (2007) and Allon et al. (2007). The first two papers define maximum likelihood estimators in semiparametric settings. The estimators in Matzkin (1991) and Banker and Maindiratta (1992) are shown to be consistent in Matzkin (1991) and Maindiratta and Sarath (1997), respectively. A maximum likelihood estimator and a sieved least squares estimator have been defined and techniques for their computation have been provided

in Allon et al. (2007) and Beresteanu (2007), respectively.

The method of least squares has been applied to multidimensional concave regression in Kuosmanen (2008). We take this work as our starting point. In agreement with the techniques used there, we define a least squares estimator which can be computed by solving a quadratic program. We argue that this estimator can be evaluated at a single point by finding the solution to a linear program. We then show that, under some mild regularity conditions, our estimator can be used to consistently estimate both, the convex function and its subdifferentials.

Our work goes beyond those mentioned above in the following ways: Our method does not require any tuning parameter(s), which is a major drawback for most nonparametric regression methods, such as kernel-based procedures. The choice of the tuning parameter(s) is especially problematic in higher dimensions, e.g., kernel based methods would require the choice of a $d \times d$ matrix of bandwidths. The sets of assumptions that most authors have used to study the estimation of a multidimensional convex regression function are more restrictive and of a different nature than the ones in this paper. As opposed to the maximum likelihood approach used in Banker and Maindiratta (1992), Matzkin (1991), Allon et al. (2007) and Maindiratta and Sarath (1997), we prove the consistency of the estimator keeping the distribution of the errors *unspecified*; e.g., in the i.i.d. case we only assume that the errors have zero expectation and finite second moment. The estimators in Beresteanu (2007) are sieved least squares estimators and assume that the observed values of the predictors lie on equidistant grids of rectangular domains. By contrast, our estimators are unsieved and our assumptions on the spatial arrangement of the predictor values are much more relaxed. In fact, we prove the consistency of the least squares estimator under both fixed and stochastic design settings; we also allow for heteroscedastic errors. In addition, we show that the least squares estimator can also be used to approximate the gradients and subdifferentials of the underlying convex function.

It is hard to overstate the importance of convex functions in applied mathematics. For instance, optimization problems with convex objective functions over convex sets appear in many applications. Thus, the question of accurately estimating a convex regression function is indeed interesting from a theoretical perspective. However, it turns out that convex regression is important for numerous reasons besides statistical curiosity. Convexity also appears in many applied sciences. One such field of application is microeconomic theory. Production functions are often supposed to be concave and componentwise nondecreasing. In this context, concavity reflects decreasing marginal returns. Concavity also plays a role in the theory of rational choice since it is a common assumption for utility functions, on which it represents decreasing marginal utility. The interested reader can see Hildreth (1954), Varian (1982a) or Varian (1982b) for more information regarding the importance of concavity/convexity in economic theory.

The paper is organized as follows. In Section 2 we discuss the estimation procedure, characterize the estimator and show how it can be computed by solving a positive semidefinite quadratic program and a linear program. Section 3 starts with a description of the deterministic and stochastic design regression schemes. The statement and proof of our main results are also included in Section 3. In Section 4 we provide the proofs of the technical lemmas used to prove the main theorem. Section A, the Appendix, contains some results from convex analysis and linear algebra that are used in the paper and may be of independent interest.

2 Characterization and finite sample properties

We start with some notation. For convenience, we will regard elements of the Euclidian space $\mathbb{R}^m$ as column vectors and denote their components with upper indices, i.e, any $z \in \mathbb{R}^m$ will be denoted as $z = (z^1, z^2, \ldots, z^m)$. The symbol $\overline{\mathbb{R}}$ will stand for the extended real line. Additionally, for any set $A \subset \mathbb{R}^d$ we will denoted as $Conv(A)$ its convex hull

and we'll write $Conv(X_1,\ldots,X_n)$ instead of $Conv(\{X_1,\ldots,X_n\})$. Finally, we will use $\langle\cdot,\cdot\rangle$ and $|\cdot|$ to denote the standard inner product and norm in Euclidian spaces, respectively.

For $\mathscr{X}=\{X_1,\ldots,X_n\}\subset\mathfrak{X}\subset\mathbb{R}^d$, consider the set $\mathscr{K}_{\mathscr{X}}$ of all vectors $z=(z^1,\ldots,z^n)'\in\mathbb{R}^n$ for which there is a convex function $\psi:\mathfrak{X}\to\mathbb{R}$ such that $\psi(X_j)=z^j$ for all $j=1,\ldots,n$. Then, a necessary and sufficient condition for a convex function ψ to minimize the sum of squared errors is that $\psi(X_j)=Z_n^j$ for $j=1,\ldots,n$, where

$$Z_n=\underset{z\in\mathscr{K}_{\mathscr{X}}}{\operatorname{argmin}}\left\{\sum_{k=1}^{n}\left|Y_k-z^k\right|^2\right\}. \tag{2}$$

The computation of the vector Z_n is crucial for the estimation procedure. We will show that such a vector exists and is unique. However, it should be noted that there are many convex functions ψ satisfying $\psi(X_j)=Z_n^j$ for all $j=1,\ldots,n$. Although any of these functions can play the role of the least squares estimator, there is one such function which is easily evaluated in $Conv(X_1,\ldots,X_n)$. For computational convenience, we will define our least squares estimator $\hat{\phi}_n$ to be precisely this function and describe it explicitly in (7) and the subsequent discussion.

In what follows we show that both, the vector Z_n and the least squares estimator $\hat{\phi}_n$ are well-defined for any n data points $(X_1,Y_1),\ldots,(X_n,Y_n)$. We will also provide two characterizations of the set $\mathscr{K}_{\mathscr{X}}$ and show that the vector Z_n can be computed by solving a positive semidefinite quadratic program. Finally, we will prove that for any $x\in Conv(X_1,\ldots,X_n)$ one can obtain $\hat{\phi}_n(x)$ by solving a linear program.

2.1 Existence and uniqueness

We start with two characterizations of the set $\mathscr{K}_{\mathscr{X}}$. The developments here are similar to those in Allon et al. (2007) and Kuosmanen (2008).

Lemma 2.1 (Primal Characterization) *Let $z=(z^1,\ldots,z^n)\in\mathbb{R}^n$. Then, $z\in\mathscr{K}_{\mathscr{X}}$ if and*

only if for every $j = 1, \ldots, n$, the following holds:

$$z^j = \inf\left\{ \sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1,\ \sum_{k=1}^n \theta^k X_k = X_j,\ \theta \geq 0,\ \theta \in \mathbb{R}^n \right\}, \tag{3}$$

where the inequality $\theta \geq 0$ holds componentwise.

Proof: Define the function $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ by

$$g(x) = \inf\left\{ \sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1,\ \sum_{k=1}^n \theta^k X_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n \right\} \tag{4}$$

where we use the convention that $\inf(\emptyset) = +\infty$. By Lemma A.1 in the Appendix, g is convex and finite on the X_j's. Hence, if z^j satisfies (3) then $z^j = g(X_j)$ for every $j = 1, \ldots, n$ and it follows that $z \in \mathcal{K}_{\mathcal{X}}$.

Conversely, assume that $z \in \mathcal{K}_{\mathcal{X}}$ and $g(X_j) \neq z^j$ for some j. Note that $g(X_k) \leq z^k$ for any k from the definition of g. Thus, we may suppose that $g(X_j) < z^j$. As $z \in \mathcal{K}_{\mathcal{X}}$, there is a convex function ψ such that $\psi(X_k) = z^k$ for all $k = 1, \ldots, n$. Then, from the definition of $g(X_j)$ there exist $\theta_0 \in \mathbb{R}^n$ with $\theta_0 \geq 0$ and $\theta_0^1 + \ldots + \theta_0^n = 1$ such that $\theta_0^1 X_1 + \ldots + \theta_0^n X_n = X_j$ and

$$\sum_{k=1}^n \theta_0^k \psi(X_k) = \sum_{k=1}^n \theta_0^k z^k < z^j = \psi(X_j) = \psi\left(\sum_{k=1}^n \theta_0^k X_k \right),$$

which leads to a contradiction because ψ is convex. □

We now provide an alternative characterization of the set $\mathcal{K}_{\mathcal{X}}$ based on the dual problem to the linear program used in Lemma 2.1.

Lemma 2.2 (Dual Characterization) *Let $z \in \mathbb{R}^n$. Then, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if for any $j = 1, \ldots, n$ we have*

$$z^j = \sup\left\{ \langle \xi, X_j \rangle + \eta : \langle \xi, X_k \rangle + \eta \leq z^k \ \ \forall\ k = 1, \ldots, n,\ \xi \in \mathbb{R}^d,\ \eta \in \mathbb{R} \right\}. \tag{5}$$

Moreover, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if there exist vectors $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ such that

$$\langle \xi_j, X_k - X_j \rangle \leq z^k - z^j \ \ \forall\ k, j \in \{1, \ldots, n\}. \tag{6}$$

Proof: According to the primal characterization, $z \in \mathcal{K}_{\mathcal{X}}$ if and only if the linear programs defined by (3) have the z^j's as optimal values. The linear programs in (5) are the dual problems to those in (3). Then, the duality theorem for linear programs (see Luenberger (1984), page 89) implies that the z^j's have to be the corresponding optimal values to the programs in (5).

To prove the second assertion let us first assume that $z \in \mathcal{K}_{\mathcal{X}}$. For each $j \in \{1,\ldots,n\}$ take any solution (ξ_j, η_j) to (5). Then by (5), $\eta_j = z^j - \langle \xi_j, X_j \rangle$ and the inequalities in (6) follow immediately because we must have $\langle \xi_j, X_k \rangle + \eta_j \leq z^k$ for any $k \in \{1,\ldots,n\}$. Conversely, take $z \in \mathbb{R}^n$ and assume that there are $\xi_1,\ldots,\xi_n \in \mathbb{R}^d$ satisfying (6). Take any $j \in \{1,\ldots,n\}$, $\eta_j = z^j - \langle \xi_j, X_j \rangle$ and θ to be the vector in $\mathbb{R}^n$ with components $\theta^k = \delta_{kj}$, where δ_{kj} is the Kronecker δ. It follows that $\langle \xi_j, X_k \rangle + \eta_j \leq z^k \ \forall \ k = 1,\ldots,n$ so (ξ_j, η_j) is feasible for the linear program in (5). In addition, θ is feasible for the linear program in (3) so the weak duality principle of linear programming (see Luenberger (1984), Lemma 1, page 89) implies that $\langle \xi, X_j \rangle + \eta \leq z^j$ for any pair (ξ, η) which is feasible for the problem in the right-hand side of (5). We thus have that z^j is an upper bound attained by the feasible pair (ξ_j, η_j) and hence (5) holds for all $j = 1,\ldots,n$. □

Both, the primal and dual characterizations are useful for our purposes. The primal plays a key role in proving the existence and uniqueness of the least squares estimator. The dual is crucial for its computation.

Lemma 2.3 *The set $\mathcal{K}_{\mathcal{X}}$ is a closed, convex cone in $\mathbb{R}^n$ and the vector Z_n satisfying (2) is uniquely defined.*

Proof: That $\mathcal{K}_{\mathcal{X}}$ is a convex cone follows trivially from the definition of the set. Now, if $z \notin \mathcal{K}_{\mathcal{X}}$, then there is $j \in \{1,\ldots,n\}$ for which $z^j > g(X_j)$ with the function g defined as in (4). Thus, there is $\theta_0 \in \mathbb{R}^n$ with $\theta_0 \geq 0$ and $\theta_0^1 + \ldots + \theta_0^n = 1$ such that $\theta_0^1 X_1 + \ldots + \theta_0^n X_n = X_j$ and $\sum_{k=1}^n \theta_0^k z^k < z^j$. Setting $\delta = \frac{1}{2}\left(z^j - \sum_{k=1}^n \theta_0^k z^k\right)$ it is easily seen that for

all $\zeta \in \prod_{k=1}^{n}(z^k - \delta, z^k + \delta)$ we still have $\sum_{k=1}^{n} \theta_0^k \zeta^k < \zeta^j$ and thus $\zeta \notin \mathcal{K}_{\mathcal{X}}$. Therefore we have shown that for any $z \notin \mathcal{K}_{\mathcal{X}}$ there is a neighborhood U of z with $U \subset \mathbb{R}^n \setminus \mathcal{K}_{\mathcal{X}}$. Therefore, $\mathcal{K}_{\mathcal{X}}$ is closed and the vector Z_n is uniquely determined as the projection of $(Y_1, \ldots, Y_n) \in \mathbb{R}^n$ onto the closed convex set $\mathcal{K}_{\mathcal{X}}$ (see Conway (1985), Theorem 2.5, page 9). □

We are now in a position to define the least squares estimator. Given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from model (1), we take the nonparametric least squares estimator to be the function $\hat{\phi}_n : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\hat{\phi}_n(x) = \inf\left\{ \sum_{k=1}^{n} \theta^k Z_n^k : \sum_{k=1}^{n} \theta^k = 1,\ \sum_{k=1}^{n} \theta^k X_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n \right\} \tag{7}$$

for any $x \in \mathbb{R}^d$. Here we are taking the convention that $\inf(\emptyset) = +\infty$. This function is well-defined because the vector Z_n exists and is unique for the sample. The estimator is, in fact, a polyhedral convex function (i.e., a convex function whose epigraph is a polyhedral; see Rockafellar (1970), page 172) and satisfies, as a consequence of Lemma A.1,

$$\hat{\phi}_n(x) = \sup_{\psi \in \mathcal{K}_{\mathcal{X}, Z_n}} \{\psi(x)\},$$

where $\mathcal{K}_{\mathcal{X}, Z_n}$ is the collection of all convex functions $\psi : \mathbb{R}^d \to \mathbb{R}$ such that $\psi(X_j) \leq Z_n^j$ for all $j = 1, \ldots, n$. Thus, $\hat{\phi}_n$ is the largest convex function that never exceeds the Z_n^j's. It is immediate that $\hat{\phi}_n$ is indeed a convex function (as the supremum of any family of convex functions is itself convex). The primal characterization of the set $\mathcal{K}_{\mathcal{X}}$ implies that $\hat{\phi}_n(X_j) = Z_n^j$ for all $j = 1, \ldots, n$.

2.2 Finite sample properties

In the following lemma we state some of the most important finite sample properties of the least squares estimator defined by (7).

Lemma 2.4 *Let $\hat{\phi}_n$ be the least squares estimator obtained from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then,*

(i) $\sum_{k=1}^{n} (\psi(X_k) - \hat{\phi}_n(X_k))(Y_k - \hat{\phi}_n(X_k)) \leq 0$ *for any convex function ψ which is finite on $Conv(X_1, \ldots, X_n)$;*

(ii) $\sum_{k=1}^{n} \hat{\phi}_n(X_k)(Y_k - \hat{\phi}_n(X_k)) = 0$;

(iii) $\sum_{k=1}^{n} Y_k = \sum_{k=1}^{n} \hat{\phi}_n(X_k)$;

(iv) the set on which $\hat{\phi}_n < \infty$ is $Conv(X_1, \ldots, X_n)$;

(v) for any $x \in \mathbb{R}^d$ the map $(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \hookrightarrow \hat{\phi}_n(x)$ is a Borel-measurable function from $\mathbb{R}^{n(d+1)}$ into $\mathbb{R}$.

Proof: Property (i) follows from Moreau's decomposition theorem, which can be stated as:

Consider a closed convex set $\mathscr{C}$ on a Hilbert space $\mathscr{H}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Then, for any $x \in \mathscr{H}$ there is only one vector $x_{\mathscr{C}} \in \mathscr{C}$ satisfying $\|x - x_{\mathscr{C}}\| = argmin_{\xi \in \mathscr{C}} \{\|x - \xi\|\}$. The vector $x_{\mathscr{C}}$ is characterized by being the only element of $\mathscr{C}$ for which the inequality $\langle \xi - x_{\mathscr{C}}, x - x_{\mathscr{C}} \rangle \leq 0$ holds for every $\xi \in \mathscr{C}$ (see Moreau (1962) or Song and Zhengjun (2004)).

Taking ψ to be $\kappa \hat{\phi}_n$ and letting κ vary through $(0, \infty)$ gives (ii) from (i). Similarly, (iii) follows from (i) by letting ψ to be $\hat{\phi}_n \pm 1$. Property (iv) is obvious from the definition of $\hat{\phi}_n$.

To see why (v) holds, we first argue that the map $(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \hookrightarrow Z_n$ is measurable. This follows from the fact that Z_n is the solution to a convex quadratic program and thus can be found as a limit of sequences whose elements come from arithmetic operations with $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. Examples of such sequences are the ones produced by active set methods, e.g, see Boland (1997); or by interior-point methods (see

Kapoor and Vaidya (1986) or Mehrotra and Sun (1990)). The measurability of $\hat{\phi}_n(x)$ follows from a similar argument, since it is the optimal value of a linear program whose solution can be obtained from arithmetic operations involving just $(X_1,\ldots,X_n,Y_1,\ldots,Y_n)$ and Z_n (e.g., via the well-known simplex method; see Nocedal and Wright (1999), page 372 or Luenberger (1984), page 30). □

2.3 Computation of the estimator

Once the vector Z_n defined in (2) has been obtained, the evaluation of $\hat{\phi}_n$ at a single point x can be carried out by solving the linear program in (7). Thus, we need to find a way to compute Z_n. And here the dual characterization proves of vital importance, since it allows us to compute Z_n by solving a quadratic program.

Lemma 2.5 *Consider the positive semidefinite quadratic program*

$$\begin{array}{ll} \min & \sum_{k=1}^n |Y_k - z^k|^2 \\ \textit{subject to} & \langle \xi_k, X_j - X_k\rangle \le z^j - z^k \ \ \forall\ k,j = 1,\ldots,n \\ & \xi_1,\ldots,\xi_n \in \mathbb{R}^d, z \in \mathbb{R}^n. \end{array} \tag{8}$$

Then, this program has a unique solution Z_n in z, i.e., for any two solutions $(\xi_1,\ldots,\xi_n,z)$ and $(\tau_1,\ldots,\tau_n,\zeta)$ we have $z=\zeta=Z_n$. This solution Z_n is the only vector in $\mathbb{R}^n$ which satisfies (2).

Proof: From Lemma 2.2 if $(\xi_1,\ldots,\xi_n,z)$ belongs in the feasible set of this program, then $z \in \mathcal{K}_{\mathcal{X}}$. Moreover, for any $z \in \mathcal{K}_{\mathcal{X}}$ there are $\xi_1,\ldots,\xi_n \in \mathbb{R}^d$ such that $(\xi_1,\ldots,\xi_n,z)$ belongs to the feasible set of the quadratic program. Since the objective function only depends on z, solving the quadratic program is the same as getting the element of $\mathcal{K}_{\mathcal{X}}$ which is the closest to Y. This element is, of course, the uniquely defined Z_n satisfying (2). □

The quadratic program (8) is positive semidefinite. This implies certain computational complexities, but most modern nonlinear programming solvers can handle this type of optimization problems. Some examples of high-performance quadratic programming solvers are CPLEX, LINDO,
MOSEK and QPOPT. Here we present two simulated examples to illustrate the compu-

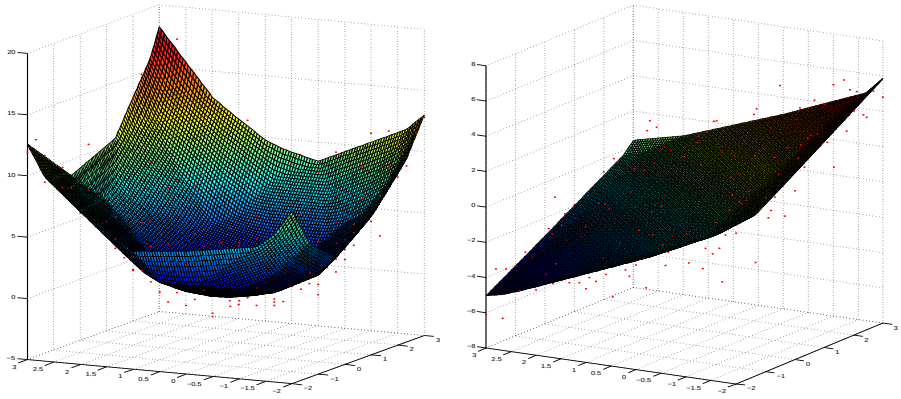

Figure 1: The scatter plot and nonparametric least squares estimator of the convex regression function when (a) $\phi(x) = |x|^2$ (left panel); (b) $\phi(x) = -x^1 + x^2$ (right panel).

tation of the estimator when $d = 2$. The first one, depicted in Figure 1a corresponds to the case where $\phi(x) = |x|^2$. Figure 1b shows the convex function estimator when the regression function is the hyperplane $\phi(x) = -x^1 + x^2$. In both cases, $n = 256$ observations were used and the errors were assumed to be i.i.d. from the standard normal distribution. All the computations were carried out using the MOSEK optimization toolbox for Matlab and the run time for each example was less than 2 minutes in a standard desktop PC. We refer the reader to Kuosmanen (2008) for additional numerical examples (although the examples there are for the estimation of concave, componentwise nondecreasing functions, the computational complexities are the same).

2.4 The componentwise nonincreasing case

We now consider the case where the regression function ϕ is assumed to be convex and componentwise nonincreasing. The developments here are quite similar to those in the convex case, so we omit some of the details. Given the observed values $(X_1, Y_1), \ldots, (X_n, Y_n)$, we write $\mathscr{Q}_{\mathscr{X}}$ for the collection of all vectors $z \in \mathbb{R}^n$ for which there is a convex, componentwise nonincreasing function ψ satisfying $\psi(X_j) = z^j$ for every $j = 1, \ldots, n$. We will denote by $\mathbb{R}^d_+$ and $\mathbb{R}^d_-$, respectively, the nonnegative and nonpositive orthants of $\mathbb{R}^d$. We now have the following characterizations.

Lemma 2.6 *Let $z \in \mathbb{R}^n$. Then, $z \in \mathscr{Q}_{\mathscr{X}}$ if and only if the following holds for every $j = 1, \ldots, n$:*

$$z^j = \inf\left\{ \sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1,\ \vartheta + \sum_{k=1}^n \theta^k X_k = X_j,\ \theta \geq 0,\ \theta \in \mathbb{R}^n, \vartheta \in \mathbb{R}^d_+ \right\}.$$

Proof: The proof is very similar to that of Lemma 2.1. The difference being that we use Lemma A.2 and the function

$$h(x) = \inf\left\{ \sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1,\ \vartheta + \sum_{k=1}^n \theta^k X_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n, \vartheta \in \mathbb{R}^d_+ \right\}$$

instead of using Lemma A.1 and the function g. □

The analogous dual characterization here is given in the following lemma. Its proof is just an application of the duality theorem of linear programming, so we omit it.

Lemma 2.7 *Let $z \in \mathbb{R}^n$. Then, $z \in \mathscr{Q}_{\mathscr{X}}$ if and only if for every $j = 1, \ldots, n$ we have*

$$z^j = \sup\left\{ \langle \xi, X_j \rangle + \eta : \langle \xi, X_k \rangle + \eta \leq z^k \ \forall\ k = 1, \ldots, n,\ \xi \in \mathbb{R}^d_-,\ \eta \in \mathbb{R} \right\}.$$

Moreover, $z \in \mathscr{Q}_{\mathscr{X}}$ if and only if there exist vectors $\xi_1, \ldots, \xi_n \in \mathbb{R}^d_-$ such that

$$\langle \xi_j, X_k - X_j \rangle \leq z^k - z^j \ \ \forall\ k, j \in \{1, \ldots, n\}.$$

Just as in the previous case, we can use both characterizations to show the existence and uniqueness of the vector

$$W_n = \operatorname*{argmin}_{z \in \mathcal{Q}_{\mathcal{X}}} \left\{ \sum_{k=1}^{n} \left| Y_k - z^k \right|^2 \right\}$$

and then define the nonparametric least squares estimator by

$$\hat{\varphi}_n(x) = \inf \left\{ \sum_{k=1}^{n} \theta^k W_n^k : \sum_{k=1}^{n} \theta^k = 1, \vartheta + \sum_{k=1}^{n} \theta^k X_k = x, \theta \in \mathbb{R}_+^n, \vartheta \in \mathbb{R}_+^d \right\}.$$

Here, the vector W_n can also be computed by solving the corresponding quadratic program

$$\begin{array}{ll} \min & \sum_{k=1}^{n} |Y_k - z^k|^2 \\ \text{subject to} & \langle \xi_k, X_j - X_k \rangle \leq z^j - z^k \ \ \forall\ k, j = 1, \ldots, n \\ & \xi_1, \ldots, \xi_n \in \mathbb{R}_-^d, z \in \mathbb{R}^n. \end{array}$$

which differs from the program (8) just because here the ξ_j's have to be nonpositive. The estimator enjoys analogous finite dimensional properties to those listed in Lemma 2.4. For the sake of completeness, we include them in the following lemma.

Lemma 2.8 *Let $\hat{\varphi}_n$ be the convex, componentwise nonincreasing least squares estimator obtained from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then,*

(i) $\sum_{k=1}^{n} (\psi(x_k) - \hat{\varphi}_n(X_k))(Y_k - \hat{\varphi}_n(X_k)) \leq 0$ *for any convex, componentwise nonincreasing function ψ which is finite on $Conv(X_1, \ldots, X_n)$;*

(ii) $\sum_{k=1}^{n} \hat{\varphi}_n(X_k)(Y_k - \hat{\varphi}_n(X_k)) = 0$;

(iii) $\sum_{k=1}^{n} Y_k = \sum_{k=1}^{n} \hat{\varphi}_n(X_k)$;

(iv) the set on which $\hat{\varphi}_n < \infty$ is $Conv(X_1, \ldots, X_n) + \mathbb{R}_+^d$;

(v) for any $x \in \mathbb{R}^d$ the map $(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \hookrightarrow \hat{\varphi}_n(x)$ is a Borel-measurable function from $\mathbb{R}^{n(d+1)}$ into $\mathbb{R}$.

3 Consistency of the least squares estimator

The main goal of this paper is to show that in an appropriate setting the nonparametric least squares estimator $\hat{\phi}_n$ described above is consistent for estimating the convex function ϕ on the set $\mathfrak{X}$. In this context, we will prove the consistency of $\hat{\phi}_n$ in both, fixed and stochastic design regression settings.

Before proceeding any further we would like to introduce some notation. For any Borel set $\mathfrak{X} \subset \mathbb{R}^d$ we will denote by $\mathscr{B}_{\mathfrak{X}}$ the σ-algebra of Borel subsets of $\mathfrak{X}$. Given a sequence of events $(A_n)_{n=1}^{\infty}$ we will be using the notation $[A_n \text{ i.o.}]$ and $[A_n \text{ a.a.}]$ to denote $\overline{\lim}\, A_n$ and $\underline{\lim}\, A_n$, respectively.

Now, consider a convex function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. This function is said to be proper if $f(x) > -\infty$ for every $x \in \mathbb{R}^d$. The effective domain of f, denoted by $\mathrm{Dom}(f)$, is the set of points $x \in \mathbb{R}^d$ for which $f(x) < \infty$. The subdifferential of f at a point $x \in \mathbb{R}^d$ is the set $\partial f(x) \subset \mathbb{R}^d$ of all vectors ξ satisfying the inequality

$$\langle \xi, h \rangle \leq f(x+h) - f(x) \quad \forall\ h \in \mathbb{R}^d.$$

The elements of $\partial f(x)$ are called subgradients of f at x (see Rockafellar (1970)). For a set $A \subset \mathbb{R}^d$ we denote by A°, $\overline{A}$ and ∂A its interior, closure and boundary, respectively. We write $\mathrm{Ext}(A) = \mathbb{R}^d \setminus \overline{A}$ for the exterior of the set A and $\mathrm{diam}(A) := \sup_{x,y \in A} |x - y|$ for the diameter of A. We also use the sup-norm notation, i.e., for a function $g : \mathbb{R}^d \to \mathbb{R}$ we write $\|g\|_A = \sup_{x \in A} |g(x)|$.

To avoid measurability issues regarding some sets, specially those involving the random set-valued functions $\{\partial \hat{\phi}_n(x)\}_{x \in \mathfrak{X}^{\circ}}$, we will use the symbols $\mathbf{P}_*$ and $\mathbf{P}^*$ to denote inner and outer probabilities, respectively. We refer the reader to Van der Vaart and Wellner (1996), pages 6-15, for the basic properties of inner and outer probabilities. In this context, a sequence of (not necessarily measurable) functions $(\Psi_n)_{n=1}^{\infty}$ from a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ into $\mathbb{R}$ is said to converge to a function Ψ *almost surely* (see

Van der Vaart and Wellner (1996), Definition 1.9.1-(iv), page 52), written $\Psi_n \xrightarrow{a.s.} \Psi$, if $\mathbf{P}_*(\Psi_n \to \Psi) = 1$. We will use the standard notation $\mathbf{P}(A)$ for the probabilities of all events A whose measurability can be easily inferred from the measurability of the random variables $\{\hat{\phi}_n(x)\}_{x\in\mathfrak{X}}$, established in Lemma 2.4.

Our main theorems hold for both, fixed and stochastic design schemes, and the proofs are very similar. They differ only in minor steps. Therefore, for the sake of simplicity, we will denote the observed values of the regressor variables always with the capital letters X_n. For any Borel set $\mathtt{X} \subset \mathbb{R}^d$, we write

$$N_n(\mathtt{X}) = \#\{1 \le j \le n : X_j \in \mathtt{X}\}.$$

The quantities X_n and $N_n(\mathtt{X})$ are non-random under the fixed design but random under the stochastic one.

3.1 Fixed Design

In a "fixed design" regression setting we assume that the regressor values are non-random and that all the uncertainty in the model comes from the response variable. We will now list a set of assumptions for this type of design. The one-dimensional case has been proven, under different regularity conditions, in Hanson and Pledger (1976).

(A1) We assume that we have a sequence $(X_n, Y_n)_{n=1}^{\infty}$ satisfying

$$Y_k = \phi(X_k) + \epsilon_k$$

where $(\epsilon_n)_{n=1}^{\infty}$ is an i.i.d. sequence with $\mathbf{E}(\epsilon_j) = 0$, $\mathbf{E}\left(\epsilon_j^2\right) = \sigma^2 < \infty$ and $\phi : \mathbb{R}^d \to \mathbb{R}$ is a proper convex function.

(A2) The non-random sequence $(X_n)_{n=1}^{\infty}$ is contained in a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$ with $\mathfrak{X}^\circ \neq \emptyset$ and $\mathfrak{X} \subset \mathrm{Dom}(\phi)$.

(A3) We assume the existence of a Borel measure ν on $\mathfrak{X}$ satisfying:

(i) $\{\mathtt{X} \in \mathscr{B}_{\mathfrak{X}} : \nu(\mathtt{X}) = 0\} = \{\mathtt{X} \in \mathscr{B}_{\mathfrak{X}} : \mathtt{X} \text{ has Lebesgue measure } 0\}$.

(ii) $\frac{1}{n} N_n(\mathtt{X}) \to \nu(\mathtt{X})$ for any open rectangle $\mathtt{X} \subset \mathfrak{X}^\circ$.

Condition (A1) may be replaced by the following:

(A4) We assume that we have a sequence $(X_n, Y_n)_{n=1}^\infty$ satisfying

$$Y_k = \phi(X_k) + \epsilon_k$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a proper convex function and $(\epsilon_n)_{n=1}^\infty$ is an independent sequence of random variables satisfying

(i) $\mathbf{E}(\epsilon_n) = 0 \ \forall \ n \in \mathbb{N}$ and $\underline{\lim} \frac{1}{n} \sum_{k=1}^n \mathbf{E}(|\epsilon_k|) > 0$.

(ii) $\sum_{n=1}^\infty \frac{\mathrm{Var}(\epsilon_n^2)}{n^2} < \infty$.

(iii) $\sup_{n \in \mathbb{N}} \{\mathbf{E}\left(\epsilon_n^2\right)\} < \infty$.

Under these conditions we define $\sigma^2 := \overline{\lim}_{n \to \infty} \frac{1}{n} \sum_{j=1}^n \mathbf{E}\left(\epsilon_j^2\right)$.

The raison d'etre of condition (A4) is to allow the variance of the error terms to depend on the regressors. We make the distinction between (A1) and (A4) because in the case of i.i.d. errors it is enough to require a finite second moment to ensure consistency.

3.2 Stochastic Design

In this setting we assume that $(X_n, Y_n)_{n=1}^\infty$ is an i.i.d. sequence from some Borel probability measure μ on $\mathbb{R}^{d+1}$. Here we make the following assumptions on the measure μ:

(A5) There is a closed, convex set $\mathfrak{X} \subset \mathbb{R}^d$ with $\mathfrak{X}^\circ \neq \emptyset$ such that $\mu(\mathfrak{X} \times \mathbb{R}) = 1$. Also,

$$\int_{\mathfrak{X} \times \mathbb{R}} y^2 \mu(dx, dy) < \infty.$$

(A6) There is a proper convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\mathfrak{X} \subset \mathrm{Dom}(\phi)$ such that whenever $(X, Y) \sim \mu$ we have $\mathbf{E}\big(Y - \phi(X)|X\big) = 0$ and $\mathbf{E}\big(|Y - \phi(X)|^2\big) = \sigma^2 < \infty$. Thus, ϕ is the regression function.

(A7) Denoting by $\nu(\cdot) = \mu((\cdot) \times \mathbb{R})$ the x-marginal of μ, we assume that

$$\{\mathtt{X} \in \mathscr{B}_{\mathfrak{X}} : \nu(\mathtt{X}) = 0\} = \{\mathtt{X} \in \mathscr{B}_{\mathfrak{X}} : \mathtt{X} \text{ has Lebesgue measure } 0\}.$$

We wish to point out some conclusions that one can draw from these assumptions. Consider the class of functions

$$\mathscr{K}_\mu := \left\{\psi : \mathbb{R}^d \to \mathbb{R} \mid \psi \text{ is convex with } \int |\psi(x)|^2 \nu(dx) < \infty\right\}.$$

Then for any $\mathtt{X} \subset \mathfrak{X}$ the following holds

$$\int_{\mathtt{X}\times\mathbb{R}} \psi(x)(y - \phi(x))\mu(dx, dy) = 0 \ \ \forall \psi \ \in \mathscr{K}_\mu;$$

so we get that ϕ is in fact the element of $\mathscr{K}_\mu$ which is the closest to Y in the Hilbert space $\mathbb{L}^2(\mathtt{X} \times \mathbb{R}, \mathscr{B}_{\mathtt{X}\times\mathbb{R}}, \mu)$. This follows from Moreau's decomposition theorem (see the proof of Lemma 2.4).

Additionally, conditions {A5-A7} allow for stochastic dependency between the error variable $Y - \phi(X)$ and the regressor X. Although some level of dependency can be put to satisfy conditions {A2-A4}, the measure μ allows us to take into account some cases which wouldn't fit in the fixed design setting (even by conditioning on the regressors).

3.3 Main results

We can now state the two main results of this paper. The first result shows that assuming only the convexity of ϕ, the least squares estimator can be used to consistently estimate both ϕ and its subdifferentials $\partial\phi(x)$.

Theorem 3.1 *Under any of {A1-A3}, {A2-A4} or {A5-A7} we have,*

(i) $\mathbf{P}\left(\sup_{x\in X}\{|\hat{\phi}_n(x)-\phi(x)|\}\to 0 \text{ for any compact set } X\subset\mathfrak{X}^\circ\right)=1.$

(ii) For every $x\in\mathfrak{X}^\circ$ *and every* $\xi\in\mathbb{R}^d$

$$\overline{\lim_{n\to\infty}}\lim_{h\downarrow 0}\frac{\hat{\phi}_n(x+h\xi)-\hat{\phi}_n(x)}{h}\le\lim_{h\downarrow 0}\frac{\phi(x+h\xi)-\phi(x)}{h} \text{ almost surely.}$$

(iii) Denoting by $\boldsymbol{B}$ *the unit ball (w.r.t. the Euclidian norm) we have*

$$\mathbf{P}_*\left(\partial\hat{\phi}_n(x)\subset\partial\phi(x)+\epsilon\boldsymbol{B}\ \text{a.a.}\right)=1\quad\forall\,\epsilon>0,\ \forall\ x\in\mathfrak{X}^\circ.$$

(iv) If ϕ *is differentiable at* $x\in\mathfrak{X}^\circ$*, then*

$$\sup_{\xi\in\partial\hat{\phi}_n(x)}\{|\xi-\nabla\phi(x)|\}\xrightarrow{a.s.}0.$$

Our second result states that assuming differentiability of ϕ on the entire $\mathfrak{X}^\circ$ allows us to use the subdifferentials of the least squares estimator to consistently estimate $\nabla\phi$ uniformly on compact subsets of $\mathfrak{X}^\circ$.

Theorem 3.2 *If* ϕ *is differentiable on* $\mathfrak{X}^\circ$*, then under any of {A1-A3}, {A2-A4} or {A5-A7} we have,*

$$\mathbf{P}_*\left(\sup_{\substack{x\in X\\ \xi\in\partial\hat{\phi}_n(x)}}\{|\xi-\nabla\phi(x)|\}\to 0 \text{ for any compact set } X\subset\mathfrak{X}^\circ\right)=1.$$

3.4 Proof of the main results

Before embarking on the proofs, one must notice that there are some statements which hold true under any of {A1-A3}, {A2-A4} or {A5-A7}. We list the most important ones below, since they'll be used later.

- For any set $\mathrm{X}\subset\mathfrak{X}$ we have

$$\frac{N_n(\mathrm{X})}{n}\xrightarrow{a.s.}\nu(\mathrm{X}). \tag{9}$$

- The strong law of large numbers implies that for any Borel set $\mathtt{X} \subset \mathfrak{X}$ with positive Lebesgue measure we have

$$\frac{1}{N_n(\mathtt{X})} \sum_{\substack{X_k \in \mathtt{X} \\ 1 \le k \le n}} (Y_k - \phi(X_k)) \xrightarrow{a.s.} 0 \tag{10}$$

and also

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{1 \le k \le n} (Y_k - \phi(X_k))^2 = \sigma^2 \text{ a.s.} \tag{11}$$

We would like to point out that in the case of condition A4, A4-(iii) allows us to obtain (10) from an application of a version of the strong law of large number for uncorrelated random variables, as it appears in Chung (2001), page 108, Theorem 5.1.2. Similarly, condition A4-(ii) implies that we can apply a version the strong law of large numbers for independent random variables as in Williams (1991), Lemma 12.8, page 118 or in Folland (1999), Theorem 10.12, page 322 to obtain (11).

- For any Borel subset $\mathtt{X} \subset \mathfrak{X}$ with positive Lebesgue measure,

$$\#\{n \in \mathbb{N} : X_n \in \mathtt{X}\} \xrightarrow{a.s.} +\infty \tag{12}$$

Proof of Theorem 3.1. We will only make distinctions among the design schemes in the proof if we are using any property besides (9), (10), (11) or (12). For the sake of clarity, we divide the proof in steps.

Step I: We start by showing that for any set with positive Lebesgue measure there is a uniform band around the regression function (over that set) such that $\hat{\phi}_n$ comes within the band at least at one point for all but finitely many n's. This fact is stated in the following lemma (proved in Section 4.1).

Lemma 3.1 *For any set $X \subset \mathfrak{X}$ with positive Lebesgue measure we have,*

$$\mathbf{P}\left(\inf_{x\in X}\{|\hat{\phi}_n(x)-\phi(x)|\} \geq M \text{ i.o.}\right) = 0 \;\; \forall \; M > \frac{\sigma}{\sqrt{\nu(X)}}.$$

Step II: The idea is now to use the convexity of both, ϕ and $\hat{\phi}_n$, to show that the previous result in fact implies that the sup-norm of $\hat{\phi}_n$ is uniformly bounded on compact subsets of $\mathfrak{X}^\circ$. We achieve this goal in the following two lemmas (whose proofs are given in Sections 4.2 and 4.3 respectively).

Lemma 3.2 *Let $X \subset \mathfrak{X}^\circ$ be compact with positive Lebesgue measure. Then, there is a positive real number K_X such that*

$$\mathbf{P}\left(\inf_{x\in X}\{\hat{\phi}_n(x)\} < -K_X \text{ i.o.}\right) = 0.$$

Lemma 3.3 *Let $X \subset \mathfrak{X}^\circ$ be a compact set with positive Lebesgue measure. Then, there is $K_X > 0$ such that*

$$\mathbf{P}\left(\sup_{x\in X}\{\hat{\phi}_n(x)\} \geq K_X \text{ i.o.}\right) = 0.$$

Step III: Convex functions are determined by their subdifferential mappings (see Rockafellar (1970), Theorem 24.9, page 239). Moreover, having a uniform upper bound $K_{\mathtt{X}}$ for the norms of all the subgradients over a compact region $\mathtt{X}$ imposes a Lipschitz continuity condition on the convex function over $\mathtt{X}$ (see Rockafellar (1970), Theorem 24.7, page 237); the Lipschitz constant being $K_{\mathtt{X}}$. For these reasons, it is important to have a uniform upper bound on the norms of the subgradients of $\hat{\phi}_n$ on compact regions. The following lemma (proved in Section 4.4) states that this can be achieved.

Lemma 3.4 *Let $X \subset \mathfrak{X}^\circ$ be a compact set with positive Lebesgue measure. Then, there is $K_X > 0$ such that*

$$\mathbf{P}^*\left(\sup_{\substack{\xi\in\partial\hat{\phi}_n(x)\\ x\in X}} \{|\xi|\} > K_X \text{ i.o.}\right) = 0.$$

Step IV: For the next results we need to introduce some further notation. We will denote by μ_n the empirical measure defined on $\mathbb{R}^{d+1}$ by the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. In agreement with Van der Vaart and Wellner (1996), given a class of functions $\mathscr{G}$ on $D \subset \mathbb{R}^{d+1}$, a seminorm $\|\cdot\|$ on some space containing $\mathscr{G}$ and $\epsilon > 0$ we denote by $N(\epsilon, \mathscr{G}, \|\cdot\|)$ the ϵ covering number of $\mathscr{G}$ with respect to $\|\cdot\|$.

Although Lemmas 3.5 and 3.7 may seem unrelated to what has been done so far, they are crucial for the further developments. Lemma 3.5 (proved in Section 4.5) shows that the class of convex functions is not very complex in terms of entropy. Lemma 3.7 is a uniform version of the strong law of large numbers which proves vital in the proof of Lemma 3.8.

Lemma 3.5 *Let $\mathcal{X} \subset \mathfrak{X}^\circ$ be a compact rectangle with positive Lebesgue measure. For $K > 0$ consider the class $\mathscr{G}_{K,\mathcal{X}}$ of all functions of the form $\psi(X)(Y - \phi(X))\mathbf{1}_{\mathcal{X}}(X)$ where ψ ranges over the class $\mathscr{D}_{K,\mathcal{X}}$ of all proper convex functions which satisfy*

(a) $\|\psi\|_{\mathcal{X}} \leq K$;

(b) $\bigcup_{\substack{\xi \in \partial\psi(x) \\ x \in \mathcal{X}}} \{\xi\} \subset [-K, K]^d$.

Then, for any $\epsilon > 0$ we have

$$\overline{\lim_{n\to\infty}} N(\epsilon, \mathscr{G}_{K,\mathcal{X}}, \mathbb{L}_1(\mathcal{X} \times \mathbb{R}, \mu_n)) < \infty \text{ almost surely,}$$

and there is a positive constant $A_\epsilon < \infty$, depending only on $(X_1, \ldots, X_n)$, K and $\mathcal{X}$, such that the covering numbers $N(\frac{\epsilon}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|, \mathscr{G}_{K,\mathcal{X}}, \mathbb{L}_1(\mathcal{X} \times \mathbb{R}, \mu_n))$ are bounded above by A_ϵ, for all $n \in \mathbb{N}$, almost surely.

The proofs of Lemmas 3.7 and 3.8 (given in Sections 4.7 and 4.8 respectively) are the only parts in the whole proof where we must treat the different design schemes separately. To make the argument work, a small lemma (proved in Section 4.6) for the set of

conditions {A2-A4} is required. We include it here for the sake of completeness and to point out the difference between the schemes.

Lemma 3.6 *Consider the set of conditions {A2-A4} and a subsequence* $(n_k)_{k=1}^{\infty}$ *such that*

$$\lim_{k\to\infty} \frac{1}{n_k} \sum_{j=1}^{n_k} \mathbf{E}\left(\epsilon_j^2\right) = \sigma^2.$$

Let $(X_m)_{m=1}^{\infty}$ *be a an increasing sequence of compact subsets of* $\mathfrak{X}$ *satisfying* $\nu(X_m) \to 1$. *Then,*

$$\lim_{m\to\infty} \varliminf_{k\to\infty} \frac{1}{n_k} \sum_{\{1\le j\le n_k : X_j \in X_m\}} \mathbf{E}\left(\epsilon_j^2\right) = \sigma^2.$$

We are now ready to state the key result on the uniform law of large numbers.

Lemma 3.7 *Consider the notation of Lemma 3.5 and let* $X \subset \mathfrak{X}^\circ$ *be any finite union of compact rectangles with positive Lebesgue measure. Then,*

$$\sup_{\psi\in\mathscr{D}_{K,X}} \left\{ \left| \frac{1}{n} \sum_{\{1\le j\le n : X_j \in X\}} \psi(X_j)(Y_j - \phi(X_j)) \right| \right\} \xrightarrow{a.s.} 0.$$

Step V: With the aid of all the results proved up to this point, it is now possible to show that Lemma 3.1 is in fact true if we replace M by an arbitrarily small $\eta > 0$. The proof of the following lemma is given in Section 4.8.

Lemma 3.8 *Let* $X \subset \mathfrak{X}^\circ$ *be any compact set with positive Lebesgue measure. Then,*

(i) $\mathbf{P}\left(\inf_{x\in X}\{\phi(x) - \hat{\phi}_n(x)\} \ge \eta \text{ i.o.}\right) = 0 \quad \forall\, \eta > 0,$

(ii) $\mathbf{P}\left(\sup_{x\in X}\{\phi(x) - \hat{\phi}_n(x)\} \le -\eta \text{ i.o.}\right) = 0 \quad \forall\, \eta > 0.$

Step VI: Combining the last lemma with the fact that we have a uniform bound on the norms of the subgradients on compacts, we can state and prove the consistency result on compacts. This is done in the next lemma (proof included in Section 4.9).

Lemma 3.9 *Let* $X \subset \mathfrak{X}^\circ$ *be a compact set with positive Lebesgue measure. Then,*

(i) $\mathbf{P}\left(\inf_{x\in X}\{\hat{\phi}_n(x)-\phi(x)\}<-\eta \text{ i.o.}\right)=0 \quad \forall\,\eta>0,$

(ii) $\mathbf{P}\left(\sup_{x\in X}\{\hat{\phi}_n(x)-\phi(x)\}>\eta \text{ i.o.}\right)=0 \quad \forall\,\eta>0,$

(iii) $\sup_{x\in X}\{|\hat{\phi}_n(x)-\phi(x)|\}\xrightarrow{a.s.}0.$

Step VII: We can now complete the proof of Theorem 3.1. Consider the class $\mathfrak{C}$ of all open rectangles $\mathscr{R}$ such that $\overline{\mathscr{R}}\subset\mathfrak{X}^\circ$ and whose vertices have rational coordinates. Then, $\mathfrak{C}$ is countable and $\bigcup_{\mathscr{R}\in\mathfrak{C}}\mathscr{R}=\mathfrak{X}^\circ$. Observe that Lemmas 3.2 and 3.3 imply that for any finite union $A:=\mathscr{R}_1\cup\cdots\cup\mathscr{R}_m$ of open rectangles $\mathscr{R}_1,\ldots,\mathscr{R}_m\in\mathfrak{C}$ there is, with probability one, $n_0\in\mathbb{N}$ such that the sequence $(\hat{\phi}_n)_{n=n_0}^{\infty}$ is finite on $Conv(A)$. From Lemma 3.9 we know that the least squares estimator converges at all rational points in $\mathfrak{X}^\circ$ with probability one. Then, Theorem 10.8, page 90 of Rockafellar (1970) implies that (i) holds if $\mathfrak{X}^\circ$ is replaced by the convex hull of a finite union of rectangles belonging to $\mathfrak{C}$. Since there are countably many of such unions and any compact subset of $\mathfrak{X}^\circ$ is contained in one of those unions, we see that (i) holds. An application of Theorem 24.5, page 233 of Rockafellar (1970) on an open rectangle C containing x and satisfying $\overline{C}\subset\mathfrak{X}^\circ$ gives (ii) and (iii). Note that (iv) is a consequence of (iii). □

Proof of Theorem 3.2. To prove the desired result we need the following lemma (whose proof is provided in Section 4.10) from convex analysis. The result is an extension of Theorem 25.7, page 248 of Rockafellar (1970), and might be of independent interest.

Lemma 3.10 *Let $\mathscr{C}\subset\mathbb{R}^d$ be an open, convex set and f a convex function which is finite and differentiable on $\mathscr{C}$. Consider a sequence of convex functions $(f_n)_{n=1}^{\infty}$ which are finite on $\mathscr{C}$ and such that $f_n\to f$ pointwise on $\mathscr{C}$. Then, if $X\subset\mathscr{C}$ is any compact set,*

$$\sup_{\substack{x\in X\\ \xi\in\partial f_n(x)}}\left\{|\xi-\nabla f(x)|\right\}\to 0.$$

Defining the class $\mathfrak{C}$ of open rectangles as in the proof of Theorem 3.1, one can use a similar argument to obtain Theorem 3.2 from an application of Theorem 3.1 and the previous lemma. □

3.5 The componentwise nonincreasing case

The regression function ϕ is now assumed to be convex and componentwise nonincreasing. Recalling the notation defined in Section 2.4, we now have that Theorems 3.1 and 3.2 still hold with $\hat{\phi}_n$ replaced by $\hat{\varphi}_n$. In view of the fact that the proof of the results is very similar to that when ϕ is just convex, we omit the proof and sketch the main differences. The proof of the main results in Section 3 relied essentially on two key facts:

(i) The finite sample properties of $\hat{\phi}_n$ established in Lemma 2.4.

(ii) The vector $(\hat{\phi}_n(X_1),\ldots,\hat{\phi}_n(X_n))' \in \mathbb{R}^n$ is the $\mathscr{L}_2$ projection of $(Y_1,\ldots,Y_n)$ on the closed, convex cone $\mathscr{K}_{\mathscr{X}}$ of all evaluations of proper convex functions on $(X_1,\ldots,X_n)$. Also, note that $(\phi(X_1),\ldots,\phi(X_n))' \in \mathscr{K}_{\mathscr{X}}$.

We know from Lemma 2.8 that $\hat{\varphi}_n$ has similar finite sample properties as its convex counterpart. Note that if ϕ is convex and componentwise nonincreasing $(\phi(X_1),\ldots,\phi(X_n))' \in \mathscr{Q}_{\mathscr{X}}$ and $(\hat{\varphi}_n(X_1),\ldots,\hat{\varphi}_n(X_n))' \in \mathbb{R}^n$ is the $\mathscr{L}_2$ projection of $(Y_1,\ldots,Y_n)$ onto $\mathscr{Q}_{\mathscr{X}}$.

From these considerations and the nature of the arguments used to prove Theorems 3.1 and 3.2, it follows that all but one of those arguments carry forward to the componentwise nonincreasing case; the only difference being the entropy calculation of Lemma 3.5. At some point in that proof, one breaks the rectangle $[-K,K]^d$ into a family of subrectangles in order to approximate the subdifferentials of the class $\mathscr{D}_{K,\mathbb{X}}$. It is easily seen that the same argument holds in the componentwise nonincreasing case

if one instead uses a partition of $[-K,0]^d$ to approach the subdifferentials of the corresponding class $\mathscr{D}_{K,\mathrm{X}}$ for componentwise nonincreasing convex functions. By doing this, the resulting function g will be convex and componentwise nonincreasing and (30), (31) and (32) will still hold for the corresponding class $\mathscr{H}_{n,c}$. Then, the conclusions of Lemma 3.5 are also true for the componentwise nonincreasing case and we can conclude that our main results are valid in this case too.

4 Proofs of the lemmas

Here we prove the lemmas involved in the proof of the main theorem. To prove these, we will need additional auxiliary results from matrix algebra and convex analysis, which may be of independent interest and are proved in the Appendix.

4.1 Proof of Lemma 3.1

We will first show that the event
$\left[\inf_{x\in\mathrm{X}}\left\{\hat{\phi}_n(x)-\phi(x)\right\}\geq M \text{ i.o.}\right]$ has probability zero. Under this event, there is a subsequence $(n_k)_{k=1}^{\infty}$ such that $\inf_{x\in\mathrm{X}}\left\{\hat{\phi}_{n_k}(x)-\phi(x)\right\}\geq M\ \forall\ k\in\mathbb{N}$. Then (10) implies that for this subsequence, with probability one, we have

$$\overline{\lim_{k\to\infty}}\frac{1}{N_{n_k}(\mathrm{X})}\sum_{X_j\in\mathrm{X}}\{Y_j-\hat{\phi}_{n_k}(X_j)\} \quleq -M. \tag{13}$$

On the other hand, it is seen (by solving the corresponding quadratic programming problems; see, e.g., Exercise 16.2, page 484 of Nocedal and Wright (1999)) that for any $\eta>0$, $m\in\mathbb{N}$

$$\inf\left\{\frac{1}{m}\sum_{1\leq j\leq m}|\xi^j|^2 : \frac{1}{m}\sum_{1\leq j\leq m}\xi^j\geq\eta,\ \xi\in\mathbb{R}^m\right\} = \eta^2, \tag{14}$$

$$\inf\left\{\frac{1}{m}\sum_{1\leq j\leq m}|\xi^j|^2 : \frac{1}{m}\sum_{1\leq j\leq m}\xi^j\leq-\eta,\ \xi\in\mathbb{R}^m\right\} = \eta^2. \tag{15}$$

For $0 < \delta < M$, using (15) with $\eta = M - \delta$ together with (12) and (13) we get that, with probability one, we must have

$$\varliminf_{k\to\infty} \frac{1}{n_k} \sum_{j=1}^{n_k} (Y_j - \hat{\phi}_{n_k}(X_j))^2 \geq \nu(\mathrm{X})(M-\delta)^2.$$

Letting $\delta \to 0$ we actually get

$$\varliminf_{k\to\infty} \frac{1}{n_k} \sum_{j=1}^{n_k} (Y_j - \hat{\phi}_{n_k}(X_j))^2 \geq \nu(\mathrm{X})M^2 > \sigma^2 = \varlimsup_{k\to\infty} \frac{1}{n_k} \sum_{j=1}^{n_k} (Y_j - \phi(X_j))^2 \text{ a.s.}$$

which is impossible because $\hat{\phi}_{n_k}$ is the least squares estimator. Therefore,

$$\mathbf{P}\left(\inf_{x\in\mathrm{X}} \{\hat{\phi}_n(x) - \phi(x)\} \geq M \text{ i.o.}\right) = 0.$$

A similar argument now using (14) gives

$$\mathbf{P}\left(\sup_{x\in\mathrm{X}} \{\hat{\phi}_n(x) - \phi(x)\} \leq -M \text{ i.o.}\right) = 0,$$

which completes the proof of the lemma. □

Before we prove Lemmas 3.2 and 3.3, we need some additional results from matrix algebra. For convenience, we state them here, but postpone their proofs to Section A.2 in the Appendix.

We first introduce some notation. We write $\mathbf{e}_j \in \mathbb{R}^d$ for the vector whose components are given by $\mathbf{e}_j^k = \delta_{jk}$, where δ_{jk} is the Kronecker δ. We also write $\mathbf{e} = \mathbf{e}_1 + \ldots + \mathbf{e}_d$ for the vector of ones in $\mathbb{R}^d$. For $\alpha \in \{-1,1\}^d$ we write

$$\mathscr{R}_\alpha = \left\{ \sum_{k=1}^{d} \theta^k \alpha^k \mathbf{e}_k : \theta \geq 0, \theta \in \mathbb{R}^d \right\}$$

for the orthant in the α direction. For any hyperplane $\mathscr{H}$ defined by the normal vector $\xi \in \mathbb{R}^d$ and the intercept $b \in \mathbb{R}$, we write $\mathscr{H} = \{x \in \mathbb{R}^d : \langle \xi, x \rangle = b\}$, $\mathscr{H}^+ = \{x \in \mathbb{R}^d : \langle \xi, x \rangle > b\}$ and $\mathscr{H}^- = \{x \in \mathbb{R}^d : \langle \xi, x \rangle < b\}$. For $r > 0$ and $x_0 \in \mathbb{R}^d$ we will write $B(x_0, r) = \{x \in \mathbb{R}^d : |x - x_0| < r\}$. We denote by $\mathbb{R}^{d\times d}$ the space of $d \times d$ matrices endowed with the topology

defined by the $\|\cdot\|_2$ norm (where $\|A\|_2 = \sup_{|x|\le 1}\{|Ax|\}$ and can be shown to be equal to the largest singular value of A; see Harville (2008)).

Lemma 4.1 *Let $r > 0$. There is a constant $R_r > 0$, depending only on r and d, such that for any $\rho_* \in (0, R_r)$ there are $\rho, \rho^* > 0$ with the property: for any $\alpha \in \{-1,1\}^d$ and any d-tuple of vectors $\beta = \{x_1, \ldots, x_d\} \subset \mathbb{R}^d$ such that $x_j \in B(\alpha^j r\mathbf{e}_j, \rho)\ \forall\ j = 1, \ldots, d$, there is a unique pair $(\xi_{\alpha,\beta}, b_{\alpha,\beta})$, with $\xi_{\alpha,\beta} \in \mathbb{R}^d$, $|\xi_{\alpha,\beta}| = 1$ and $b_{\alpha,\beta} > 0$ for which the following statements hold:*

(i) *β form a basis for $\mathbb{R}^d$.*

(ii) *$x_1, \ldots, x_d \in \mathcal{H}_{\alpha,\beta} := \{x \in \mathbb{R}^d : \langle \xi_{\alpha,\beta}, x\rangle = b_{\alpha,\beta}\}$.*

(iii) *$\min\limits_{1\le j\le d}\{|\xi^j_{\alpha,\beta}|\} > 0$.*

(iv) *$B(0,\rho_*) \subset \mathcal{H}^-_{\alpha,\beta}$.*

(v) *$\{x \in \mathbb{R}^d : |x| \ge \rho^*\} \cap \mathcal{R}_\alpha \subset \mathcal{H}^+_{\alpha,\beta}$.*

(vi) *$B(-\alpha^j r\mathbf{e}_j, \rho) \subset \{x \in \mathbb{R}^d : \langle \xi_{\alpha,\beta}, x\rangle < 0\}$ for all $j = 1, \ldots, d$.*

(vii) *For any $w_1 \in B\left(0, \frac{\rho_*}{16\sqrt{d}}\right)$ and $w_2 \in B\left(\frac{3\rho_*}{8\sqrt{d}}\alpha, \frac{\rho_*}{8\sqrt{d}}\right)$ we have*

$$\min_{1\le j\le d}\left\{\left(X_\beta^{-1}(w_1 + t(w_2 - w_1))\right)^j\right\} > 0\ \ \forall\ t \ge 1$$

where $X_\beta = (x_1, \ldots, x_d) \in \mathbb{R}^{d\times d}$ is the matrix whose j'th column is x_j.

Figure 2a illustrates the above lemma when $d = 2$ and $\alpha = (1,1)$. The lemma states that whatever points x_1 and x_2 are taken inside the circles of radius ρ around $\alpha^1 r\mathbf{e}_1$ and $\alpha^2 r\mathbf{e}_2$, respectively, $B(0,\rho_*)$ and $\{x \in \mathbb{R}^d : |x| \ge \rho^*\} \cap \mathcal{R}_\alpha$ are contained, respectively, in the half-spaces $\mathcal{H}^-_{\alpha,\beta}$ and $\mathcal{H}^+_{\alpha,\beta}$. Assertion (vii) of the lemma implies that all the points in the half line $\{w_1 + t(w_2 - w_1\}_{t\ge 1}$ should have positive co-ordinates with respect to the basis β as they do with respect to the basis $\{\alpha^j\mathbf{e}_j\}_{j=1}^d$. We refer the reader to Section A.2.1 for a complete proof of Lemma 4.1.

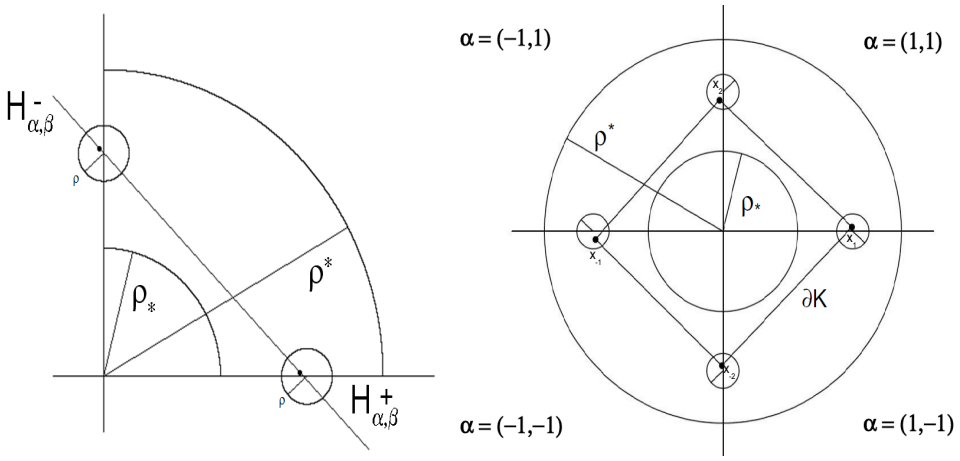

Figure 2: Explanatory diagram for (a) Lemma 4.1 (left panel); (b) Lemma 4.2 (right panel).

We now state two other useful results, namely Lemma 4.2 and Lemma 4.3, but defer their proofs to Section A.2.2 and Section A.2.3 respectively.

Lemma 4.2 *Let $r > 0$ and consider the notation of Lemma 4.1 with the positive numbers ρ, ρ_* and ρ^* as defined there. Take 2d vectors $\{x_{\pm 1}, \ldots, x_{\pm d}\} \subset \mathbb{R}^d$ such that $x_{\pm j} \in B(\pm r\boldsymbol{e}_j, \rho)$ and for $\alpha \in \{-1,1\}^d$ write $\beta_\alpha = \{x_{\alpha^1 1}, x_{\alpha^2 2}, \ldots, x_{\alpha^d d}\}$, $\xi_\alpha = \xi_{\alpha,\beta_\alpha}$, $b_\alpha = b_{\alpha,\beta_\alpha}$ and $\mathscr{H}_\alpha = \mathscr{H}_{\alpha,\beta}$, all in agreement with the setting of Lemma 4.1. Then, if $K = Conv\,(x_{\pm 1}, \ldots, x_{\pm d})$ we have:*

(i) $K = \bigcap_{\alpha\in\{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x\rangle \le b_\alpha\}$.

(ii) $K^\circ = \bigcap_{\alpha\in\{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x\rangle < b_\alpha\}$.

(iii) $\partial K = \bigcup_{\alpha\in\{-1,1\}^d} Conv\left(x_{\alpha^1 1}, \ldots, x_{\alpha^d d}\right)$.

(iv) $\partial K = \left(\bigcup_{\alpha\in\{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x\rangle = b_\alpha\}\right) \bigcap \left(\bigcap_{\alpha\in\{-1,1\}^d} \{x \in \mathbb{R}^d : \langle \xi_\alpha, x\rangle \le b_\alpha\}\right)$.

(v) $B(0, \rho_*) \subset K^\circ$.

(vi) $\partial B(0, \rho^*) \subset \mathrm{Ext}(K)$.

Figure 2b illustrates Lemma 4.2 for the two-dimensional case. Intuitively, the idea is that as long as the points $x_{\pm 1}$ and $x_{\pm 2}$ belong to $B(\pm r\mathbf{e}_1, \rho)$ and $B(\pm r\mathbf{e}_2, \rho)$, respectively, we will have $B(0, \rho_*)$ and $\partial B(0, \rho^*)$ as subsets of K° and $\text{Ext}(K)$, respectively.

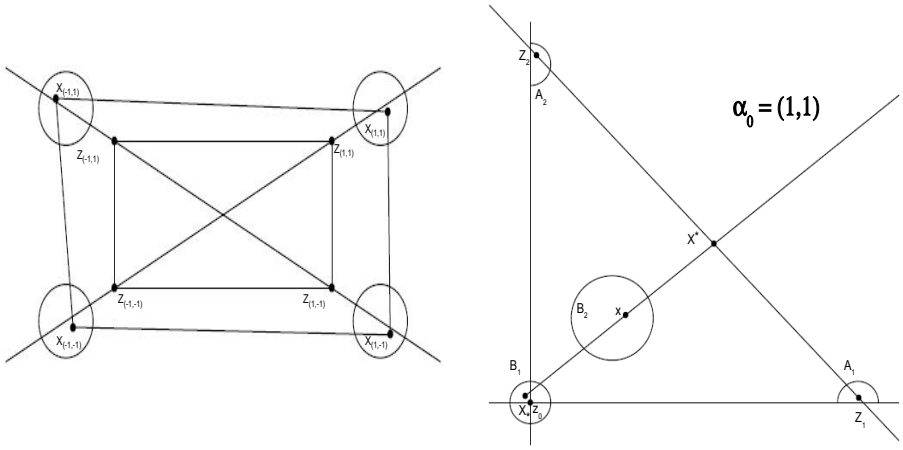

Figure 3: Explanatory diagram for (a) Lemma 4.3 (left panel); (b) Lemma 3.2 (right panel).

Lemma 4.3 *Let $[a,b] \subset \mathbb{R}^d$ be a compact rectangle and $r > 0$, with $r < \frac{1}{d-2}$ if $d \geq 3$. For each $\alpha \in \{-1,1\}^d$ write $z_\alpha = a + \sum_{j=1}^d \frac{1+\alpha^j}{2}\left(b^j - a^j\right)\boldsymbol{e}_j$ so that $\{z_\alpha\}_{\alpha\in\{-1,1\}^d}$ is the set of vertices of $[a,b]$. Then, there is $\rho > 0$ such that if $x_\alpha \in B(z_\alpha + r(z_\alpha - z_{-\alpha}), \rho)\ \forall\ \alpha \in \{-1,1\}^d$, then*

$$[a,b] \subset Conv\left(x_\alpha : \alpha \in \{-1,1\}^d\right)^\circ.$$

Figure 3a describes Lemma 4.3 in the two-dimensional case. As long as the points $x_{(\pm 1,\pm 1)}$ are chosen in the balls of radius ρ around $z_{(\pm 1,\pm 1)} + r(z_{(\pm 1,\pm 1)} - z_{(\mp 1,\mp 1)})$, $Conv\left(x_{(\pm 1,\pm 1)}\right)$ will contain $Conv\left(z_{(\pm 1,\pm 1)}\right)$.

4.2 Proof of Lemma 3.2

Since any compact subset of $\mathfrak{X}^\circ$ is contained in a finite union of compact rectangles, it is enough to prove the result when $\mathtt{X}$ is a compact rectangle $[a,b] \subset \mathfrak{X}^\circ$. Let $r =$

$\frac{1}{4}\min_{1\le k\le d}\{b^k-a^k\}$ and choose $\rho\in(0,\frac{1}{4}r)$, $\rho^*>0$ and $0<\rho_*<\frac{1}{2}r$ such that the conclusions of Lemmas 4.1 and 4.2 hold for any $\alpha\in\{-1,1\}^d$ and any $\beta=(z_1,\ldots,z_d)\in\mathbb{R}^{d\times d}$ with $z_j\in B(\alpha^j r\mathbf{e}_j,\rho)$. Take $N\in\mathbb{N}$ such that

$$\frac{1}{N}\max_{1\le k\le d}\{b^k-a^k\}<\frac{1}{32d}\rho_* \tag{16}$$

and divide $\mathtt{X}$ into N^d rectangles all of which are geometrically identical to $\frac{1}{N}[0,b-a]$. Let $\mathscr{C}$ be any one of the rectangles in the grid and choose any vertex z_0 of $\mathscr{C}$ satisfying

$$z_0=\operatorname*{argmax}_{z\in\mathscr{C}}\left\{\max_{1\le j\le d}\left\{z^j-a^j,b^j-z^j\right\}\right\}.$$

Then, from the definition of z_0 and r, there is $\alpha_0\in\{-1,1\}^d$ such that

$$B(z_0,r)\cap\left(z_0+\mathscr{R}_{\alpha_0}\right)\subset\mathtt{X}.$$

Additionally, define

$$\begin{aligned}
B_1 &= B\left(z_0,\frac{\rho_*}{16\sqrt{d}}\right),\\
B_2 &= B\left(z_0+\frac{3\rho_*}{8\sqrt{d}}\alpha_0,\frac{\rho_*}{8\sqrt{d}}\right),\\
A_j &= B(z_0+\alpha_0^j r\mathbf{e}_j,\rho)\cap(z_0+\mathscr{R}_{\alpha_0})\quad\forall\ j=1,\ldots,d,\\
A_{-j} &= B(z_0-\alpha_0^j r\mathbf{e}_j,\rho)\quad\forall\ j=1,\ldots,d.
\end{aligned}$$

Observe that all the sets in the previous display have positive Lebesgue measure and that the A_{-j}'s are not necessarily contained in $\mathtt{X}$. Let $M_1=\|\phi\|_{\mathtt{X}}$, $M_0>\frac{\sigma}{\sqrt{\min\{\nu(B_1),\nu(B_2),\nu(A_1),\ldots,\nu(A_d)\}}}$, $M=M_1+M_0$ and $K_{\mathscr{C}}>6M$. Also, notice that $\mathscr{C}\subset B_1$ because of (16). We will argue that

$$\mathbf{P}\left(\inf_{x\in\mathscr{C}}\{\hat{\phi}_n(x)\}\le -K_{\mathscr{C}}\text{ i.o.}\right)=0. \tag{17}$$

From Lemma 3.1, we know that

$$\mathbf{P}\left(\bigcap_{j=1}^{d}\left[\inf_{x\in A_j}\left\{\left|\hat{\phi}_n(x)-\phi(x)\right|\right\}<M_0\text{ a.a.}\right]\right)=1, \tag{18}$$

so there is, with probability one, $n_0 \in \mathbb{N}$ such that $\inf_{x\in A_j}\{|\hat{\phi}_n(x)-\phi(x)|\} < M_0$ for any $n \geq n_0$ and any $j = 1,\ldots,d$.

Assume that the event $\left[\inf_{x\in\mathscr{C}}\{\hat{\phi}_n(x)\} < -K_{\mathscr{C}} \text{ i.o.}\right]$ is true. Then, there is a subsequence n_k such that $\inf_{x\in\mathscr{C}}\{\hat{\phi}_{n_k}(x)\} < -K_{\mathscr{C}}$ for all $k \in \mathbb{N}$. Fix any $k \geq n_0$. We know that there is $X_* \in \mathscr{C} \subset B_1$ such that $\hat{\phi}_{n_k}(X_*) \leq -K_{\mathscr{C}}$. In addition, for $j = 1,\ldots,d$, there are $Z_{\alpha_0^j j} \in A_j$ such that $|\hat{\phi}_{n_k}(Z_{\alpha_0^j j}) - \phi(Z_{\alpha_0^j j})| < M_0$, which in turn implies $\hat{\phi}_{n_k}(Z_{\alpha_0^j j}) < M$. Pick any $Z_{-\alpha_0^j} \in A_{-j}$ and let $K = Conv\,(Z_{\pm 1},\ldots,Z_{\pm d}) = z_0 + Conv\,(Z_{\pm 1}-z_0,\ldots,Z_{\pm d}-z_0)$.

Take any $x \in B_2$. We will show the existence of $X^* \in Conv\left(Z_{\alpha_0^1 1},\ldots,Z_{\alpha_0^d d}\right)$ such that $x \in Conv\,(X_*, X^*)$, as shown in Figure 3b for the case $d = 2$. We will then show that the existence of such an X^* implies that

$$|\phi(x) - \hat{\phi}_{n_k}(x)| > M_0. \tag{19}$$

Consequently, since x is an arbitrary element of B_2 we will have

$$\left[\inf_{x\in\mathscr{C}}\{\hat{\phi}_n(x)\} \leq -K_{\mathscr{C}} \text{ i.o.}\right] \cap \left(\bigcap_{j=1}^{d}\left[\inf_{x\in A_j}\{|\hat{\phi}_n(x)-\phi(x)|\} < M_0 \text{ a.a.}\right]\right)$$

$$\subset \left[\inf_{x\in B_2}\{|\phi(x)-\hat{\phi}_{n_k}(x)|\} \geq M_0 \text{ i.o.}\right].$$

But from Lemma 3.1, the event on the right is a null set. Taking (18) into account, we will see that (17) holds and then complete the argument by taking $K_{\mathfrak{X}} = \max_{\mathscr{C}}\{K_{\mathscr{C}}\}$.

To show the existence of X^* consider the function $\psi : \mathbb{R} \to \mathbb{R}^d$ given by $\psi(t) = X_* + t(x - X_*)$. The function ψ is clearly continuous and satisfies $\psi(0) = X_*$ and $\psi(1) = x \in B_2 \subset K^\circ$. That $B_2 \subset K^\circ$ is a consequence of Lemma 4.1, (iv). The set K is bounded, so there is $T > 1$ such that $\psi(T) \in \text{Ext}(K) = \mathbb{R}^d \setminus \overline{K}$. The intermediate value theorem then implies that there is $t^* \in (1,T)$ such that $X^* := \psi(t^*) \in \partial K$. Observe that by Lemma 4.2 (iii) we have

$$\partial K = \bigcup_{\alpha\in\{-1,1\}^d} Conv\left(Z_{\alpha^1 1},\ldots,Z_{\alpha^d d}\right).$$

Lemma 4.1 (i) implies that $\{Z_{\alpha_0^1 1} - z_0, \ldots, Z_{\alpha_0^d d} - z_0\}$ forms a basis of $\mathbb{R}^d$ so we can write $X^* - z_0 = \sum_{j=1}^d \theta^j (Z_{\alpha_0^j j} - z_0)$. Moreover, Lemma 4.1 (vii) implies that $\theta^j > 0$ for every $j = 1, \ldots, d$ as $\theta = (\theta^1, \ldots, \theta^d) = (Z_{\alpha_0^1 1} - z_0, \ldots, Z_{\alpha_0^d d} - z_0)^{-1}(X^* - z_0)$. Here we apply Lemma 4.1 (vii) with $w_1 = X_* \in B_1$, $w_2 = x \in B_2$ and $t^* > 1$.

For $\alpha \in \{-1, 1\}^d$ consider the pair $(\xi_\alpha, b_\alpha) \in \mathbb{R}^d \times \mathbb{R}$ as defined in Lemma 4.2 for the set of vectors $\{Z_{\pm 1} - z_0, \ldots, Z_{\pm d} - z_0\}$ (here we move the origin to z_0). Observe that Lemma 4.1 (ii) implies that $\langle \xi_{\alpha_0}, Z_{\alpha_0^j j} - z_0 \rangle = b_{\alpha_0}$ for all $j = 1, \ldots, d$. Consequently, $\langle \xi_{\alpha_0}, X^* - z_0 \rangle = b_{\alpha_0} \sum_{j=1}^d \theta^j$, but since $X^* \in \partial K$, Lemma 4.2 (iv) implies that $\langle \xi_{\alpha_0}, X^* - z_0 \rangle \leq b_{\alpha_0}$ and hence $\sum_{j=1}^d \theta^j \leq 1$. Additionally, for $\alpha \neq \alpha_0$ we can write $\langle \xi_\alpha, X^* - z_0 \rangle$ as

$$\sum_{j=1}^d \theta^j \langle \xi_\alpha, Z_{\alpha_0^j j} - z_0 \rangle = \sum_{\alpha^j = \alpha_0^j} \theta^j b_\alpha + \sum_{\alpha^j \neq \alpha_0^j} \theta^j \langle \xi_\alpha, Z_{\alpha_0^j j} - z_0 \rangle < b_\alpha \tag{20}$$

as $\langle \xi_\alpha, Z_{\alpha^j} - z_0 \rangle = b_\alpha$ (by Lemma 4.1 (ii)) and $\langle \xi_\alpha, Z_{-\alpha^j} - z_0 \rangle < 0$ (by Lemma 4.1 (vi)) for every $j = 1, \ldots, d$. Since $\langle \xi_\alpha, w - z_0 \rangle = b_\alpha$ for all $w \in Conv\left(Z_{\alpha^1 1}, \ldots, Z_{\alpha^d d}\right)$ and all $\alpha \in \{-1.1\}^d$, (20) and the fact that $X^* \in \partial K$ imply that $X^* \in Conv\left(Z_{\alpha_0^1 1}, \ldots, Z_{\alpha_0^d d}\right)$. Hence $\hat{\phi}_n(X^*) \leq \sum_{j=1}^d \theta^j \hat{\phi}_{n_k}(Z_{\alpha_0^j j}) < M$. We therefore have

$$\hat{\phi}_{n_k}(X^*) < M \quad , \quad \hat{\phi}_{n_k}(X_*) < -K_{\mathscr{C}}, \tag{21}$$

$$X_* + \frac{1}{t^*}(X^* - X_*) \quad = \quad x. \tag{22}$$

Since $X_* \in B_1$ and $d \geq 1$ we have

$$|z_0 - X_*| < \frac{1}{8}\rho_*. \tag{23}$$

By using the triangle inequality we get the following bounds

$$\frac{1}{4}\rho_* < |z_0 - x| < \frac{1}{2}\rho_*. \tag{24}$$

And from Lemma 4.1 (iv) and the fact that $\langle \xi_{\alpha_0}, X^* \rangle = b_{\alpha_0}$ we also obtain

$$|z_0 - X^*| \geq \rho_*. \tag{25}$$

From (22) we know that $t^* = \frac{|X^*-X_*|}{|x-X_*|}$. Using the triangle inequality with (23), (24) and (25) one can find lower and upper bounds for $|X^*-X_*|$ (as $|X^*-X_*| \geq |X^*-z_0|-|z_0-X_*|$) and $|x-X_*|$ (as $|x-X_*| \leq |x-z_0|+|z_0-X_*|$), respectively, to obtain $t^* \geq \frac{7}{5}$. Then, (21) and (22) imply

$$\hat{\phi}_{n_k}(x) \leq \left(1-\frac{1}{t^*}\right)\hat{\phi}_{n_k}(X_*) + \frac{1}{t^*}\hat{\phi}_{n_k}(X^*) \leq -\frac{2}{7}K_{\mathscr{C}} + \frac{5}{7}M < -M.$$

Consequently,

$$|\phi(x) - \hat{\phi}_{n_k}(x)| > M - M_1 = M_0.$$

This proves (19) and completes the proof. □

4.3 Proof of Lemma 3.3

Assume without loss of generality that $\mathtt{X}$ is a compact rectangle. Let $\{z_\alpha : \alpha \in \{-1,1\}^d\}$ be the set of vertices of the rectangle. Then, there is $r \in (0,1)$ such that $B(z_\alpha, r) \subset \mathfrak{X}^\circ \ \forall$ $\alpha \in \{-1,1\}^d$. Recall that from Lemma 4.3, there is $0 < \rho < \frac{1}{2}r$ such that for any $\{\eta_\alpha : \alpha \in \{-1,1\}^d\}$ if $\eta_\alpha \in B(z_\alpha + \frac{r}{2}(z_\alpha - z_{-\alpha}), \rho)$ then $\mathtt{X} \subset Conv\left(\eta_\alpha : \alpha \in \{-1,1\}^d\right)$.

Let $A_\alpha = B(z_\alpha + \frac{1}{2}r(z_\alpha - z_{-\alpha}), \frac{\rho}{2})$ and $M_0 > \frac{\sigma}{\sqrt{\min\{v(A_\alpha):\alpha\in\{-1,1\}^d\}}}$ and choose

$$M_1 = \sup_{x \in Conv\left(\bigcup_{\alpha\in\{-1,1\}^d} A_\alpha\right)} \{|\phi(x)|\}.$$

Take $K_{\mathtt{X}} > M_0 + M_1$. Since

$$\mathbf{P}\left(\bigcap_{\alpha\in\{-1,1\}^d}\left[\inf_{x\in A_\alpha}\{|\hat{\phi}_n(x)-\phi(x)|\} < M_0, \text{ a.a.}\right]\right) = 1$$

by Lemma 3.1, there is, with probability one, $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$ we can find $\eta_\alpha \in A_\alpha$, $\alpha \in \{-1,1\}^d$, such that $|\hat{\phi}_n(\eta_\alpha) - \phi(\eta_\alpha)| < M_0$. It follows that $\hat{\phi}_n(\eta_\alpha) \leq K_{\mathtt{X}} \ \forall \ \alpha \in \{-1,1\}^d$. Now, using Lemma 4.3 we have $\mathtt{X} \subset Conv\left(\eta_\alpha : \alpha \in \{-1,1\}^d\right)$ and the convexity of $\hat{\phi}_n$ implies that $\hat{\phi}_n(x) \leq K_{\mathtt{X}}$ for any $x \in \mathtt{X}$. □

4.4 Proof of Lemma 3.4

Assume that $\mathtt{X} = [a,b]$ is a rectangle with vertices $\{z_\alpha : \alpha \in \{-1,1\}^d\}$. The function $\psi(x) = \inf_{\eta \in \overline{\mathrm{Ext}(\mathfrak{X})}}\{|x-\eta|\}$ is continuous on $\mathbb{R}^d$ so there is $x_* \in \partial\mathtt{X}$ such that $\psi(x_*) = \inf_{x\in\partial\mathtt{X}}\{\psi(x)\}$. Observe that $\psi(x_*) > 0$ because $x_* \in \partial\mathtt{X} \subset \mathfrak{X}^\circ$. By Lemma 4.3, there is a $r < \frac{1}{2}\psi(x_*)$ for which there exists $\rho < \frac{1}{4}r$ such that whenever $\eta_\alpha \in A_\alpha := B\left(z_\alpha + \frac{3}{4}r\left(\frac{z_\alpha - z_{-\alpha}}{|z_\alpha - z_{-\alpha}|}\right), \rho\right)$ for any $\alpha \in \{-1,1\}^d$ and

$$\begin{aligned} K_z &= Conv\left(z_\alpha + \frac{1}{2}r\left(\frac{z_\alpha - z_{-\alpha}}{|z_\alpha - z_{-\alpha}|}\right) : \alpha \in \{-1,1\}^d\right) \\ K_\eta &= Conv\left(\eta_\alpha : \alpha \in \{-1,1\}^d\right) \end{aligned}$$

we have

$$\mathtt{X} \subset K_z \subset K_\eta^\circ \subset K_\eta \subset \mathfrak{X}^\circ. \tag{26}$$

Let $M_0 > \frac{\sigma}{\sqrt{\min\{\nu(A_\alpha):\alpha\in\{-1,1\}^d\}}}$ and $M_1 \in \mathbb{R}$ be such that

$$\mathbf{P}\left(\inf_{x\in\mathtt{X}}\{\hat{\phi}_n(x)\} \le -M_0 \text{ i.o.}\right) = 0 \quad \text{and} \quad M_1 = \sup_{x\in Conv\left(\bigcup_{\alpha\in\{-1,1\}^d} A_\alpha\right)}\{\phi(x)\}.$$

From Lemmas 3.1 and 3.2 we can find, with probability one, $n_0 \in \mathbb{N}$ such that $\inf_{x\in\mathtt{X}}\{\hat{\phi}_n(x)\} > -M_0$ and $\inf_{x\in A_\alpha}\{|\hat{\phi}_n(x) - \phi(x)|\} < M_0$ for any $n \ge n_0$. Define

$$\begin{aligned} M &= M_1 + M_0 \\ K_{\mathtt{X}} &= \frac{4|b-a|}{r\min_{1\le j\le d}\{b^j - a^j\}}M \end{aligned}$$

and take any $n \ge n_0$. Then, for any $\alpha \in \{-1,1\}^d$ we can find $\eta_\alpha \in A_\alpha$ such that $|\hat{\phi}_n(\eta_\alpha) - \phi(\eta_\alpha)| < M_0$. Then, (26) implies that $\hat{\phi}_n(x) \le M \ \forall x \in \mathtt{X}$. Take then $x \in \mathtt{X}$ and $\xi \in \partial\hat{\phi}_n(x)$. A connectedness argument, like the one used in the proof of Lemma 3.2, implies that there is $t_* > 0$ such that $x + t_*\xi \in \partial K_\eta$. But then we must have $t_* > \frac{r\min_{1\le j\le d}\{b^j - a^j\}}{2|\xi||b-a|}$ as a consequence of (26), since the smallest distance between ∂K_z and $\partial\mathtt{X}$ is $\frac{r\min_{1\le j\le d}\{b^j-a^j\}}{2|b-a|}$

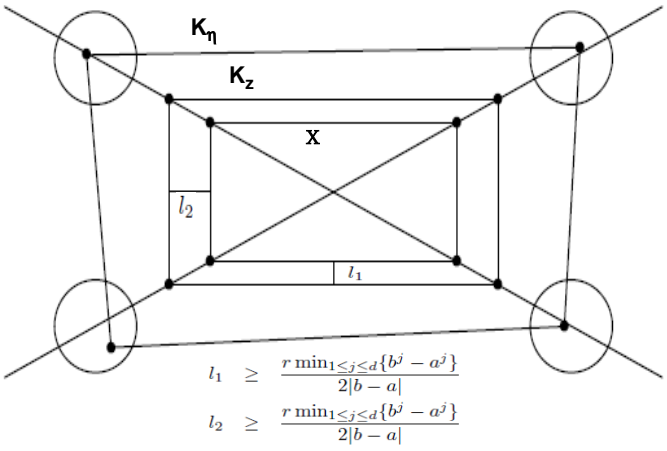

Figure 4: The smallest distance between ∂K_z and $\partial \mathbb{X}$ is at least $\frac{r \min_{1\leq j\leq d}\{b^j - a^j\}}{2|b-a|}$.

and $\partial K_\eta \subset \text{Ext}(K_z)$. This can be seen by taking a look at Figure 4, which shows the situation in the two dimensional case. Thus, using the definition of subgradients,

$$\frac{r \min_{1\leq j\leq d}\{b^j - a^j\}}{2|\xi||b-a|}\langle \xi, \xi\rangle \leq \langle \xi, t_*\xi\rangle \leq \hat{\phi}_n(x + t_*\xi) - \hat{\phi}_n(x) \leq 2M$$

which in turn implies $|\xi| \leq K_{\mathbb{X}}$. We have therefore shown that, with probability one, we can find $n_0 \in \mathbb{N}$ such that $|\xi| \leq K_{\mathbb{X}}\ \forall\ \xi \in \partial\hat{\phi}_n(x),\ \forall\ x \in \mathbb{X},\ \forall\ n \geq n_0$. This completes the proof. □

4.5 Proof of Lemma 3.5

The result is obvious for conditions {A1-A3} and {A5-A7} when $\sigma^2 = 0$. So we assume that $\sigma^2 > 0$ for {A1-A3} and {A5-A7}. Let $\epsilon > 0$ and $M = \sup_{x\in\mathbb{X}}\{|x|\}$. Choose $\delta > 0$ satisfying

$$\frac{\epsilon}{\frac{2(2M+K\sqrt{d}+1)}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|} < \delta < \frac{\epsilon}{\frac{(2M+K\sqrt{d}+1)}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|} \tag{27}$$

for n large. Notice that δ is well-defined and the quantity on the left is positive, finite and bounded away from 0 as $\underline{\lim}\,\frac{1}{n}\sum_{j=1}^n |Y_j - \phi(X_j)| > 0$ a.s. under any set of regular-

ity conditions (for {A2-A4}, conditions A4-(i) and A4-(iii) imply that we can apply the version of the strong law of large number for uncorrelated random variables, as it appears in Chung (2001), page 108, Theorem 5.1.2 to the sequence $(|\epsilon_j|)_{j=1}^{\infty}$; for {A1-A3} and {A5-A7} this is immediate as $\sigma^2 > 0$). The definition of the class $\mathscr{D}_{K,\mathtt{X}}$ implies that all its members are Lipschitz functions with Lipschitz constant bounded by $K\sqrt{d}$, a consequence of Rockafellar (1970), Theorem 24.7, page 237. Hence, (27) implies that

$$\sup_{\substack{|x-y|<\delta \\ x,y\in\mathtt{X},\psi\in\mathscr{D}_{K,\mathtt{X}}}} \{|\psi(x)-\psi(y)|\} \leq \frac{\epsilon}{\frac{1}{n}\sum_{j=1}^{n}|Y_j-\phi(X_j)|}.$$

Now, define $N_n \in \mathbb{N}$ by $N_n = \left\lceil \frac{\mathrm{diam}(\mathtt{X})}{\delta} \right\rceil \vee \left\lceil \frac{2K\sqrt{d}}{\delta} \right\rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function. Observe that (27) implies

$$N_n - 1 \leq \left(\mathrm{diam}(\mathtt{X}) \vee 2K\sqrt{d}\right) \frac{2(2M+K\sqrt{d}+1)}{\epsilon} \left(\frac{1}{n}\sum_{j=1}^{n}|Y_j-\phi(X_j)|\right). \tag{28}$$

Then, we can divide the rectangles $\mathtt{X}$ and $[-K,K]^d$ in N_n^d subrectangles, all of which have diameters less than δ. In other words, we can write

$$\begin{aligned} [-K,K]^d &= \bigcup_{1\leq j\leq N_n^d} R_j \\ \mathtt{X} &= \bigcup_{1\leq j\leq N_n^d} V_j \end{aligned}$$

with $\mathrm{diam}(R_j) < \delta$ and $\mathrm{diam}(V_j) < \delta \ \forall\ j = 1,\ldots N_n^d$. In the same way, we can divide the interval $[-K,K]$ in N_n subintervals $\mathscr{I}_1,\ldots,\mathscr{I}_{N_n}$ each having length less than δ. For each $j = 1,\ldots,N_n^d$, let ξ_j and x_j be the centroids of R_j and V_j respectively and for $j = 1,\ldots,N_n$ let η_j be the midpoint of $\mathscr{I}_j$. Consider the class of functions $\mathscr{H}_{n,\epsilon}$ defined by

$$\mathscr{H}_{n,\epsilon} = \left\{ \max_{(s,t,j)\in\mathscr{S}} \{\langle \xi_s, \cdot - x_t\rangle + \eta_j\} : \mathscr{S} \subset \{1,\ldots,N_n^d\}^2 \times \{1,\ldots,N_n\} \right\}.$$

Observe that the number of elements in the class $\mathscr{H}_{n,\epsilon}$ is bounded from above by $2^{N_n^{2d+1}}$. Now, take any $\psi \in \mathscr{D}_{K,\mathtt{X}}$. Pick any $\Xi_j \in \partial\psi(X_j)$. Then, for any j such that $X_j \in \mathtt{X}$, there

are $s_j, t_j \in \{1,\ldots,N_n^d\}$ and $\tau_j \in \{1,\ldots,N_n\}$ such that $|\Xi_j - \xi_{s_j}|$, $|X_j - x_{t_j}|$ and $|\psi(x_{t_j}) - \eta_{\tau_j}|$ are all less than δ. We then have that

$$\begin{aligned} & \sup_{x\in\mathtt{X}}\left\{\left|\langle\xi_{s_j}, x - x_{t_j}\rangle + \eta_{\tau_j} - \left(\langle\Xi_j, x - X_j\rangle + \psi(X_j)\right)\right|\right\} \\ \leq\ & 2M|\xi_{s_j} - \Xi_j| + K\sqrt{d}|x_{t_j} - X_j| + \delta < (2M + K\sqrt{d} + 1)\delta \end{aligned} \tag{29}$$

by an application of the Cauchy-Schwarz inequality. But then, (27) implies that if we define the functions $\tilde{\psi}$ and g as

$$\begin{aligned} \tilde{\psi}(x) &= \max_{X_j\in\mathtt{X}}\{\langle\Xi_j, x - X_j\rangle + \psi(X_j)\}, \\ g(x) &= \max_{X_j\in\mathtt{X}}\{\langle\xi_{s_j}, x - x_{t_j}\rangle + \eta_{\tau_j}\} \end{aligned}$$

then we have

$$\tilde{\psi}(X_j) = \psi(X_j) \text{ for } j \text{ such that } X_j \in \mathtt{X}, \tag{30}$$

$$\|g - \tilde{\psi}\|_{\mathtt{X}} \leq \frac{\epsilon}{\frac{1}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|} \text{ (from (29))}, \tag{31}$$

$$g \in \mathscr{H}_{n,\epsilon}. \tag{32}$$

Note that (30) follows from the definition of subgradients. All these facts put together give that for any $f(x,y) = \psi(x)(y - \phi(x)) \in \mathscr{G}_{K,\mathtt{X}}$, $\psi \in \mathscr{D}_{K,\mathtt{X}}$ there is $g \in \mathscr{H}_{n,\epsilon}$ such that

$$\int_{\mathtt{X}} |f(x,y) - g(x)(y - \phi(x))|\mu_n(dx, dy) < \epsilon$$

and hence

$$N(\epsilon, \mathscr{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X}\times\mathbb{R}, \mu_n)) \leq \#\mathscr{H}_{n,\epsilon} \leq 2^{N_n^{2d+1}}.$$

But then, the strong law of large numbers and (28) give that $\overline{\lim} N_n < \infty$ a.s. Furthermore, by replacing ϵ with $\frac{\epsilon}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|$ in the entire construction just made, we can see that the covering numbers $N\left(\frac{\epsilon}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|, \mathscr{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X}\times\mathbb{R}, \mu_n)\right)$ depend neither on the Y's nor on ϕ. Taking $B_\epsilon = \left(\mathrm{diam}(\mathtt{X}) \vee K\sqrt{d}\right)\frac{2(2M+K\sqrt{d}+1)}{\epsilon} + 1$ and $A_\epsilon = 2^{B_\epsilon^{2d+1}}$ it is seen that the second part of the result holds. □

4.6 Proof of Lemma 3.6

Note that for every m, we have

$$\frac{1}{n_k}\sum_{1\le j\le n_k}\mathbf{E}\left(\epsilon_j^2\right)\le\frac{1}{n_k}\sum_{\substack{X_j\in\mathtt{X}_m\\1\le j\le n_k}}\mathbf{E}\left(\epsilon_j^2\right)+\frac{N_{n_k}(\mathfrak{X}\setminus\mathtt{X}_m)}{n_k}\sup_{j\in\mathbb{N}}\{\mathbf{E}\left(\epsilon_j^2\right)\}.$$

Taking limit inferior on both sides as $k\to\infty$, we get

$$\sigma^2\le\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{\substack{X_j\in\mathtt{X}_m\\1\le j\le n_k}}\mathbf{E}\left(\epsilon_j^2\right)+\nu(\mathfrak{X}\setminus\mathtt{X}_m)\sup_{j\in\mathbb{N}}\{\mathbf{E}\left(\epsilon_j^2\right)\}.$$

Now taking the limit as $m\to\infty$ we get the result because the opposite inequality is trivial. □

4.7 Proof of Lemma 3.7

We may assume that $\mathtt{X}$ is a compact rectangle. Here we need to make a distinction between the design schemes. In the case of the stochastic design, the proof is an immediate consequence of Lemma 3.5 and Theorem 2.4.3, page 123 of Van der Vaart and Wellner (1996). Thus, we focus on the fixed design scenario.

For notational convenience, we write $M=\sup_{j\in\mathbb{N}}\{\mathbf{E}\left(\epsilon_j^2\right)\}$ and $\sum_{X_j\in\mathtt{X}}$ instead of the more cumbersome $\sum_{1\le j\le n:X_j\in\mathtt{X}}$. Letting $\epsilon_j=Y_j-\phi(X_j)$ (and using the same notation as in the proof of Lemma 3.7) first observe that the random quantity

$$\sup_{\psi\in\mathscr{D}_{K,\mathtt{X}}}\left\{\left|\frac{1}{n}\sum_{\{X_j\in\mathtt{X}\}}\psi(X_j)\epsilon_j\right|\right\}=\sup_{m\in\mathbb{N}}\left\{\sup_{g\in\mathscr{H}_{n,\frac{1}{m}}}\left\{\left|\frac{1}{n}\sum_{\{X_j\in\mathtt{X}\}}g(X_j)\epsilon_j\right|\right\}\right\}.$$

by (30), (31) and (32) and is thus measurable.

All of the following arguments are valid for both, {A1-A3} and {A2-A4}. Lyapunov's inequality (which states that for any random variable X and $1\le p\le q\le\infty$ we have $\|X\|_p\le\|X\|_q$) and the strong law of large numbers imply

$$\varlimsup_{m\to\infty}\frac{1}{m}\sum_{1\le j\le m}|\epsilon_j|=\varlimsup_{m\to\infty}\frac{1}{m}\sum_{1\le j\le m}\mathbf{E}\left(|\epsilon_j|\right)\le\sqrt{M}\text{ a.s.}\tag{33}$$

Let $\eta > 0$. From Lemma 3.5 we know that the covering numbers $a_n := N\left(\frac{\eta}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|, \mathscr{G}_{K,\mathtt{X}}, \mathbb{L}_1(\mathtt{X}\times\mathbb{R}, \mu_n)\right)$ are not random and uniformly bounded by a constant A_η. Therefore, for any $n \in \mathbb{N}$ we can find a class $\mathscr{A}_n \subset \mathscr{D}_{K,\mathtt{X}}$ with exactly a_n elements such that $\{\psi(x)(y-\phi(x))\}_{\psi\in\mathscr{A}_n}$ forms an $\left(\frac{\eta}{n}\sum_{j=1}^n |Y_j - \phi(X_j)|\right)$-net for $\mathscr{G}_{K,\mathtt{X}}$ with respect to $\mathbb{L}_1(\mathtt{X}\times\mathbb{R}, \mu_n)$. It follows that

$$\sup_{\psi\in\mathscr{D}_{K,\mathtt{X}}} \left\{ \left| \frac{1}{n} \sum_{X_j\in\mathtt{X}} \psi(X_j)\epsilon_j \right| \right\} \le \frac{\eta}{n} \sum_{1\le j\le n} |\epsilon_j| + \sup_{\psi\in\mathscr{A}_n} \left\{ \left| \frac{1}{n} \sum_{X_j\in\mathtt{X}} \psi(X_j)\epsilon_j \right| \right\}. \tag{34}$$

With (34) in mind, we make the following definitions

$$\begin{aligned}
B_n &= \sup_{\psi\in\mathscr{A}_n} \left\{ \left| \frac{1}{n} \sum_{X_j\in\mathtt{X}} \psi(X_j)\epsilon_j \right| \right\}, \\
C_n &= \sup_{\psi\in\mathscr{A}_n} \left\{ \left| \frac{1}{n} \sum_{1\le j\le \lfloor\sqrt{n}\rfloor^2:\ X_j\in\mathtt{X}} \psi(X_j)\epsilon_j \right| \right\}, \\
D_n &= \sup_{\substack{\psi\in\mathscr{A}_k \\ n^2\le k<(n+1)^2}} \left\{ \left| \frac{1}{k} \sum_{n^2<j\le k:\ X_j\in\mathtt{X}} \psi(X_j)\epsilon_j \right| \right\},
\end{aligned}$$

where $\lfloor\cdot\rfloor$ denotes the floor function. Now, pick $\delta > 0$ and observe that

$$\begin{aligned}
\mathbf{P}(B_n > \delta) &= \mathbf{P}\left(\bigcup_{\psi\in\mathscr{A}_n} \left[\left| \sum_{X_j\in\mathtt{X}} \psi(X_j)\epsilon_j \right| > n\delta \right] \right) \\
&\le \sum_{\psi\in\mathscr{A}_n} \frac{1}{n^2\delta^2} M \sum_{X_j\in\mathtt{X}} \psi(X_j)^2 \ \le\ \frac{K^2 M A_\eta}{n\delta^2}.
\end{aligned}$$

The Borel-Cantelli Lemma then implies that $\mathbf{P}\left(B_{n^2} > \delta \text{ i.o.}\right) = 0$. Letting $\delta \to 0$ through a decreasing sequence gives

$$B_{n^2} \xrightarrow{a.s.} 0. \tag{35}$$

On the other hand, the definition of C_n implies that

$$C_n \le \frac{\lfloor\sqrt{n}\rfloor^2}{n} B_{\lfloor\sqrt{n}\rfloor^2} + \frac{\eta}{n} \sum_{1\le j\le\lfloor\sqrt{n}\rfloor^2} |\epsilon_j| \tag{36}$$

which together with (35) and (33) gives

$$\overline{\lim}\, C_n \le \eta\sqrt{M} \ \text{ almost surely.} \tag{37}$$

Note that (36) is a consequence of the fact that for any $\psi \in \mathscr{A}_n$, there exists $g \in \mathscr{A}_{\lfloor\sqrt{n}\rfloor^2}$ such that if $\mathscr{J}_n = \{1 \le j \le \lfloor\sqrt{n}\rfloor^2 : X_j \in \mathtt{X}\}$, then

$$\begin{aligned}\left|\frac{1}{n}\sum_{j\in\mathscr{J}_n}\psi(X_j)\epsilon_j\right| &\le \left|\frac{1}{n}\sum_{j\in\mathscr{J}_n}(\psi(X_j)-g(X_j))\epsilon_j\right| + \left|\frac{1}{n}\sum_{j\in\mathscr{J}_n} g(X_j)\epsilon_j\right| \\ &\le \left(\frac{\lfloor\sqrt{n}\rfloor^2}{n}\right)\frac{\eta}{\lfloor\sqrt{n}\rfloor^2}\sum_{1\le j\le\lfloor\sqrt{n}\rfloor^2}|\epsilon_j| + \frac{\lfloor\sqrt{n}\rfloor^2}{n}B_{\lfloor\sqrt{n}\rfloor^2}.\end{aligned}$$

Now, a similar argument to the one used in (35) gives

$$\begin{aligned}\mathbf{P}(D_n > \delta) &= \mathbf{P}\left(\bigcup_{\substack{\psi\in\mathscr{A}_k\\ n^2\le k<(n+1)^2}}\left[\left|\sum_{n^2<j\le k:X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right| > k\delta\right]\right) \\ &\le \sum_{\substack{\psi\in\mathscr{A}_k\\ n^2\le k<(n+1)^2}}\mathbf{P}\left(\left|\sum_{n^2<j\le k:X_j\in\mathtt{X}}\psi(X_j)\epsilon_j\right| > k\delta\right) \\ &\le \sum_{\substack{\psi\in\mathscr{A}_k\\ n^2\le k<(n+1)^2}}\frac{K^2M(k-n^2)}{k^2\delta^2} \le \frac{K^2MA_\eta(2n+1)^2}{n^4\delta^2}. \end{aligned} \tag{38}$$

Again, one can use (38) and the Borel-Cantelli Lemma to prove that $\mathbf{P}(D_n > \delta \text{ i.o.}) = 0$ and then let $\delta \to 0$ through a decreasing sequence to obtain

$$D_n \xrightarrow{a.s.} 0. \tag{39}$$

Finally, one sees that

$$\sup_{\psi\in\mathscr{A}_n}\left\{\left|\frac{1}{n}\sum_{X_j\in\mathtt{X}}\psi(X_j)(Y_j-\phi(X_j))\right|\right\} = B_n \le C_n + D_{\lfloor\sqrt{n}\rfloor},$$

which combined with (37) and (39) gives

$$\overline{\lim}\, B_n \le \eta\sqrt{M} \text{ almost surely.}$$

Taking (34) into account we get

$$\overline{\lim_{n\to\infty}}\sup_{\psi\in\mathscr{D}_{K,\mathtt{X}}}\left\{\left|\frac{1}{n}\sum_{1\le j\le n:X_j\in\mathtt{X}}\psi(X_j)(Y_j-\phi(X_j))\right|\right\} \le 2\eta\sqrt{M} \text{ almost surely.}$$

Letting $\eta \to 0$ we get the desired result. □

4.8 Proof of Lemma 3.8

We can assume, without loss of generality, that $\mathtt{X}$ is a finite union of compact rectangles. Consider a sequence $(\mathtt{X}_m)_{m=1}^{\infty}$ satisfying the following properties:

(a) $\mathtt{X} \subset \mathtt{X}_m \subset \mathfrak{X}^\circ \ \forall\ m \in \mathbb{N}$.

(b) $\nu(\mathtt{X}_m) > 1 - \frac{1}{m} \ \forall\ m \in \mathbb{N}$.

(c) $\mathtt{X}_m \subset \mathtt{X}_{m+1} \ \forall\ m \in \mathbb{N}$.

(d) Every $\mathtt{X}_m$ can be expressed as a finite union of compact rectangles with positive Lebesgue measure.

The existence of such a sequence follows from the inner regularity of Borel probability measures on $\mathbb{R}^d$ and from the fact that since $\mathfrak{X}^\circ$ is open, for any compact set $F \subset \mathfrak{X}^\circ$ we can find a finite cover composed by compact rectangles with positive Lebesgue measure and completely contained in $\mathfrak{X}^\circ$. Also, from Lemmas 3.2, 3.3 and 3.4 and the fact that $\mathfrak{X} \subset \mathrm{Dom}(\phi)$, for any $m \in \mathbb{N}$ we can find $K_m > 0$ such that

$$\|\phi\|_{\mathtt{X}_m} \leq K_m \quad \text{and} \quad \mathbf{P}\left(\|\hat{\phi}_n\|_{\mathtt{X}_m} > K_m \ \text{ i.o.}\right) = 0; \tag{40}$$

$$\sup_{\substack{x \in \mathtt{X}_m \\ \xi \in \partial\phi(x)}} \{|\xi|\} \leq K_m \quad \text{and} \quad \mathbf{P}^*\left(\sup_{\substack{x \in \mathtt{X}_m \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} > K_m \ \text{ i.o.}\right) = 0. \tag{41}$$

Fix $\eta > 0$ and consider the sets

$$\begin{aligned} A &= \left[\inf_{x \in \mathtt{X}} \{\phi(x) - \hat{\phi}_n(x)\} \geq \eta \ \text{ i.o.}\right] \\ B &= \left[\|\hat{\phi}_n\|_{\mathtt{X}_m} \leq K_m \ \text{ a.a.}\right] \\ C &= \left[\sup_{\substack{x \in \mathtt{X}_m \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} \leq K_m \ \text{ a.a.}\right]. \end{aligned}$$

Suppose now that $A \cap B \cap C$ is known to be true. Then, there is a subsequence $(n_k)_{k=1}^{\infty}$ such that $\inf_{x \in \mathtt{X}}\{\phi(x) - \hat{\phi}_{n_k}(x)\} \geq \eta \ \forall\ k \in \mathbb{N}$ and $\frac{1}{n_k}\sum_{j=1}^{n_k} \mathbf{E}\left(\epsilon_j^2\right) \to \sigma^2$. Taking (40) and (41)

into account, we have that for k large enough the inequality

$$\frac{1}{n_k}\sum_{j=1}^{n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \frac{1}{n_k}\sum_{X_j\in \mathtt{X}_m}(Y_j-\phi(X_j))^2$$
$$+\frac{2}{n_k}\sum_{X_j\in \mathtt{X}_m}(Y_j-\phi(X_j))(\phi(X_j)-\hat{\phi}_{n_k}(X_j))+\frac{1}{n_k}\sum_{X_j\in \mathtt{X}_m}(\phi(X_j)-\hat{\phi}_{n_k}(X_j))^2$$

implies

$$\frac{1}{n_k}\sum_{j=1}^{n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \frac{1}{n_k}\sum_{X_j\in \mathtt{X}_m}(Y_j-\phi(X_j))^2 +$$
$$\frac{N_{n_k}(\mathtt{X})}{n_k}\eta^2-4\sup_{\psi\in\mathscr{D}_{K_m,\mathtt{X}_m}}\left\{\left|\frac{1}{n_k}\sum_{\{1\leq j\leq n_k:X_j\in\mathtt{X}_m\}}\psi(X_j)(Y_j-\phi(X_j))\right|\right\}.$$

Thus, from Lemma 3.7 we can conclude that

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{1\leq j\leq n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \nu(\mathtt{X}_m)\sigma^2+\nu(\mathtt{X})\eta^2 \text{ if } \{\text{A1-A3}\} \text{ hold.}$$

Under {A2-A4} and {A5-A7} the left-hand side of the last display is bounded from below by

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{X_j\in \mathtt{X}_m}(Y_j-\phi(X_j))^2+\nu(\mathtt{X})\eta^2$$

and

$$\int_{\mathtt{X}_m}(y-\phi(x))^2\mu(dx,dy)+\nu(\mathtt{X})\eta^2,$$

respectively.

Finally, using (a)-(d), the strong law of large numbers (for {A2-A4} we can apply a version of the strong law of large numbers for independent random variables thanks to condition A4-(ii); see Williams (1991), Lemma 12.8, page 118 or Folland (1999), Theorem 10.12, page 322) and Lemma 3.6 we can let $m\to\infty$ to see that, under any of {A1-A3}, {A2-A4} or {A5-A7},

$$\varliminf_{k\to\infty}\frac{1}{n_k}\sum_{1\leq j\leq n_k}(Y_j-\hat{\phi}_{n_k}(X_j))^2 \geq \sigma^2+\nu(\mathtt{X})\eta^2$$

which is impossible because $\hat{\phi}_{n_k}$ is the least squares estimator.

Therefore $\mathbf{P}^*(A \cap B \cap C) = 0$ and, since $\mathbf{P}_*(B \cap C) = 1$,

$$\mathbf{P}(A) = \mathbf{P}\left(\inf_{x \in \mathtt{X}} \{\phi(x) - \hat{\phi}_n(x)\} \geq \eta \text{ i.o.}\right) = 0.$$

This finishes the proof of (i). The second assertion follows from similar arguments. □

4.9 Proof of Lemma 3.9

We can assume, without loss of generality, that $\mathtt{X}$ is a finite union of compact rectangles. Pick $K_{\mathtt{X}}$ such that

$$\sup_{\substack{x \in \mathtt{X} \\ \xi \in \partial\phi(x)}} \{|\xi|\} \leq K_{\mathtt{X}} \text{ and } \mathbf{P}^*\left(\sup_{\substack{x \in \mathtt{X} \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} > K_{\mathtt{X}} \text{ i.o.}\right) = 0.$$

Let $\eta > 0$ and $\delta = \frac{\eta}{3K_{\mathtt{X}}}$. We can then divide $\mathtt{X}$ in M subrectangles $\{\mathscr{C}_1, \ldots, \mathscr{C}_M\}$ all having diameter less than δ. Define the events

$$\begin{aligned} A &= \left[\bigcap_{1 \leq k \leq M} \inf_{x \in \mathscr{C}_k} \{\hat{\phi}_n(x) - \phi(x)\} < \frac{\eta}{3} \text{ a.a.}\right] \\ B &= \left[\sup_{\substack{x \in \mathtt{X} \\ \xi \in \partial\hat{\phi}_n(x)}} \{|\xi|\} \leq K_{\mathtt{X}} \text{ a.a.}\right]. \end{aligned}$$

We will show that $A \cap B \subset \left[\sup_{x \in \mathtt{X}} \{\hat{\phi}_n(x) - \phi(x)\} \leq \eta \text{ a.a.}\right]$. Suppose $A \cap B$ is true. Then, there is $N \in \mathbb{N}$ such that for any $n \geq N$ we can find $\Xi_{n,k} \in \mathscr{C}_k$ such that $\hat{\phi}_n(\Xi_{n,k}) - \phi(\Xi_{n,k}) < \frac{\eta}{3}$. Moreover, we can make N large enough such that for any $n \geq N$, $K_{\mathtt{X}}$ is an upper bound for all the subgradients of $\hat{\phi}_n$ on $\mathtt{X}$. Then, for any $\xi \in \mathscr{C}_k$ we obtain from the Lipschitz property,

$$\begin{aligned} \hat{\phi}_n(\xi) - \phi(\xi) &= (\hat{\phi}_n(\Xi_{n,k}) - \phi(\Xi_{n,k})) + (\phi(\Xi_{n,k}) - \phi(\xi)) + (\hat{\phi}_n(\xi) - \hat{\phi}_n(\Xi_{n,k})) \\ &\leq \frac{\eta}{3} + K_{\mathtt{X}}\delta + K_{\mathtt{X}}\delta \leq \eta. \end{aligned}$$

Therefore,

$$\sup_{x \in \mathscr{C}_k} \{\hat{\phi}_n(x) - \phi(x)\} \leq \eta \quad \forall\ 1 \leq k \leq M\ \forall\ n \geq N$$

which implies

$$\sup_{x\in\mathtt{X}}\{\hat{\phi}_n(x)-\phi(x)\}\leq\eta\quad\forall\ n\geq N.$$

Considering Lemmas 3.8-(ii) and 3.4; $A\cap B\subset\left[\sup_{x\in\mathtt{X}}\{\hat{\phi}_n(x)-\phi(x)\}\leq\eta\ \text{ a.a.}\right]$ and $\mathbf{P}_*(A\cap B)=1$ we obtain (ii). The first assertion follows from similar arguments and (iii) is a direct consequence of (i) and (ii). □

4.10 Proof of Lemma 3.10

Throughout this proof we will denote by $\mathbf{B}$ the unit ball (w.r.t. the euclidian norm) in $\mathbb{R}^d$. From Theorem 25.5, page 246 on Rockafellar (1970) we know that f is continuously differentiable on $\mathscr{C}$. Let

$$h_*=\inf_{\xi\in\mathtt{X},\eta\in\mathbb{R}^d\setminus\mathscr{C}}\{|\xi-\eta|\}>0.$$

Pick $\epsilon>0$. We will first show that there is $n_\epsilon\in\mathbb{N}$ such that

$$\langle\xi,\eta\rangle\leq\langle\nabla f(x),\eta\rangle+\epsilon,\quad\forall\ \xi\in\partial f_n(x),\ \forall\ x\in\mathtt{X},\ \forall\ \eta\in\mathbf{B},\ \forall\ n\geq n_\epsilon. \tag{42}$$

Suppose that such an n_ϵ does not exist. Then, there is an increasing sequence $(m_n)_{n=1}^\infty$ such that for any $n\in\mathbb{N}$ we can find $x_{m_n}\in\mathtt{X}$, $\xi_{m_n}\in\partial f_{m_n}(x_{m_n})$, $\eta_{m_n}\in\mathbf{B}$ satisfying $\langle\xi_{m_n},\eta_{m_n}\rangle>\langle\nabla f(x_{m_n}),\eta_{m_n}\rangle+\epsilon$. But $\mathtt{X}$ and $\mathbf{B}$ are both compact, so there are $x_*\in\mathtt{X}$, $\eta_*\in\mathbf{B}$ and a subsequence $(k_n)_{n=1}^\infty$ of $(m_n)_{n=1}^\infty$ such that $x_{k_n}\to x_*$ and $\eta_{k_n}\to\eta_*$. Then, for any $0<h<h_*$ we have

$$\frac{f_{k_n}(x_{k_n}+h\eta_{k_n})-f_{k_n}(x_{k_n})}{h}\geq\langle\xi_{k_n},\eta_{k_n}\rangle>\langle\nabla f(x_{m_n}),\eta_{k_n}\rangle+\epsilon\ \ \forall\ n\in\mathbb{N},$$

and therefore

$$\lim_{n\to\infty}\lim_{h\downarrow0}\frac{f_{k_n}(x_{k_n}+h\eta_{k_n})-f_{k_n}(x_{k_n})}{h}\geq\langle\nabla f(x_*),\eta_*\rangle+\epsilon.$$

But this is impossible in view of Theorem 24.5, page 233 on Rockafellar (1970). It follows that we can choose some $n_\epsilon\in\mathbb{N}$ with the property described in (42). By noting that

$-\mathbf{B} = \mathbf{B}$, we can conclude from (42) that

$$|\langle \xi, \eta \rangle - \langle \nabla f(x), \eta \rangle| \leq \epsilon \quad \forall\ \xi \in \partial f_n(x),\ \forall\ x \in \mathbb{X},\ \forall\ \eta\ \in \mathbf{B},\ \forall\ n \geq n_\epsilon.$$

By taking $\eta_\xi = \frac{\xi - \nabla f(x)}{|\xi - \nabla f(x)|}$ when $\xi \neq \nabla f(x)$ we get

$$\sup_{\substack{x \in \mathbb{X} \\ \xi \in \partial f_n(x)}} \left\{ |\xi - \nabla f(x)| \right\} \leq \epsilon \quad \forall\ n \geq n_\epsilon.$$

Since $\epsilon > 0$ was arbitrarily chosen, this completes the proof. □

A Appendix

A.1 Results from convex analysis

Lemma A.1 *Let $z \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathbb{R}^d$ and define the function $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ by*

$$g(x) = \inf \left\{ \sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1,\ \sum_{k=1}^n \theta^k x_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n \right\}.$$

Then, g defines a convex function whose effective domain is $Conv\,(x_1, \ldots, x_n)$. Moreover, if $\mathcal{K}_{x,z}$ is the collection of all proper convex functions ψ such that $\psi(x_j) \leq z^j$ for all $j = 1, \ldots, n$, then $g = \sup_{\psi \in \mathcal{K}_{x,z}} \{\psi\}$.

Proof: To see that g defines a convex function, for any $x \in \mathbb{R}^d$ write

$$A_x = \left\{ \theta \in \mathbb{R}^n : \sum_{k=1}^n \theta^k = 1,\ \sum_{k=1}^n \theta^k x_k = x,\ \theta \geq 0 \right\}$$

and observe that for any $x, y \in \mathbb{R}^d$, $t \in (0,1)$, $\vartheta \in A_y$ and $\theta \in A_x$ we have $t\theta + (1-t)\vartheta \in A_{tx+(1-t)y}$ and hence

$$\frac{g\left(tx + (1-t)y\right) - (1-t)\sum_{k=1}^n \vartheta^k z^k}{t} \leq \sum_{k=1}^n \theta^k z^k.$$

Taking infimum over A_x and rearranging terms, we get

$$\frac{g\left(tx + (1-t)y\right) - tg(x)}{1-t} \leq \sum_{k=1}^n \vartheta^k z^k$$

and taking now the infimum over A_y gives the desired convexity. The convention that $\inf(\emptyset) = +\infty$ shows that the effective domain is precisely the convex hull of $x_1, \ldots, x_n$. Finally, for any $\psi \in \mathcal{K}_{x,z}$ and $x \in Conv(x_1, \ldots, x_n)$ we have, for $\theta \in \mathbb{R}^n$ with $\theta \geq 0$, $x = \sum_{j=1}^n \theta^j x_j$ and $\sum_{j=1}^n \theta^j = 1$,

$$\psi(x) \leq \sum_{j=1}^n \theta^j \psi(x_j) \leq \sum_{j=1}^n \theta^j z^j$$

since $\psi(x_j) \leq z^j$ for any $j = 1, \ldots, n$. The definition of g as an infimum then implies that $\psi(x) \leq g(x)\ \forall\ \psi \in \mathcal{K}_{x,z}$, $x \in Conv(x_1, \ldots, x_n)$. The result then follows from the fact that $g \in \mathcal{K}_{x,z}$. □

Lemma A.2 *Let $z \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathbb{R}^d$ and define the function $h : \mathbb{R}^d \to \overline{\mathbb{R}}$ by*

$$h(x) = \inf\left\{ \sum_{k=1}^n \theta^k z^k : \sum_{k=1}^n \theta^k = 1,\ \vartheta + \sum_{k=1}^n \theta^k X_k = x,\ \theta \geq 0,\ \theta \in \mathbb{R}^n, \vartheta \in \mathbb{R}^d_+ \right\}$$

Then, h defines a convex, componentwise nonincreasing function whose effective domain is $Conv(x_1, \ldots, x_n) + \mathbb{R}^d_+$. Moreover, if $\mathcal{Q}_{x,z}$ is the collection of all componentwise nonincreasing, proper convex functions ψ such that $\psi(x_j) \leq z^j$ for all $j = 1, \ldots, n$, then $h = \sup_{\psi \in \mathcal{Q}_{x,z}} \{\psi\}$.

Proof: The proof that h is convex is similar to the proof that g is convex in Lemma A.1. Now, if $x \leq y \in \mathbb{R}^d$, observe that for any $\theta \in \mathbb{R}^n$, $\vartheta \in \mathbb{R}^d_+$ with $\sum_{k=1}^n \theta^k = 1$, $\vartheta + \sum_{k=1}^n \theta^k X_k = x$, $\theta \geq 0$, we also have $\vartheta + (y - x) + \sum_{k=1}^n \theta^k X_k = y$ and $\vartheta + (y - x) \in \mathbb{R}^d_+$. Then, from the definition of h we see that $h(x) \geq h(y)$. Thus, h is componentwise nonincreasing. That the effective domain of h is $Conv(x_1, \ldots, x_n) + \mathbb{R}^d_+$ is clear from the fact that for any x not belonging to that set, the infimum defining $h(x)$ would be taken over the empty set. Finally, for any $\psi \in \mathcal{Q}_{x,z}$ and $x \in Conv(x_1, \ldots, x_n) + \mathbb{R}^d_+$ we have, for $\theta \in \mathbb{R}^n$ and $\vartheta \in \mathbb{R}^d_+$ with $\theta \geq 0$, $x = \vartheta + \sum_{j=1}^n \theta^j x_j$ and $\sum_{j=1}^n \theta^j = 1$,

$$\psi(x) \leq \psi\left(\sum_{j=1}^n \theta^j x_j \right) \leq \sum_{j=1}^n \theta^j \psi(x_j) \leq \sum_{j=1}^n \theta^j z^j$$

since $\psi(x_j) \leq z^j$ for any $j = 1, \ldots, n$. The definition of h as an infimum then implies that $\psi(x) \leq h(x)\ \forall\ \psi \in \mathcal{Q}_{x,z}$, $x \in Conv(x_1, \ldots, x_n) + \mathbb{R}^d_+$. The result then follows from the fact that $h \in \mathcal{Q}_{x,z}$. □

A.2 Results from matrix algebra

Before proving Lemma 4.1, we need the following result.

Lemma A.3 *Let $j \in \{1, \ldots, d\}$, $\alpha \in \{-1, 1\}^d$ and $\rho_* > 0$. Then, the optimal value of the optimization problem*

$$\begin{array}{ll} min & \langle \alpha^j \boldsymbol{e}_j, w_2 - w_1 \rangle \\ s.t. & \left| w_2 - \frac{3\rho_*}{8\sqrt{d}}\alpha \right| \leq \frac{\rho_*}{8\sqrt{d}} \\ & |w_1| \leq \frac{\rho_*}{16\sqrt{d}} \\ & w_1, w_2 \in \mathbb{R}^d \end{array}$$

is $\frac{3}{16\sqrt{d}}\rho_*$ and it is attained at $w_1^* = \frac{\rho_*}{16\sqrt{d}}\alpha^j \boldsymbol{e}_j$ and $w_2^* = \frac{3\rho_*}{8\sqrt{d}}\alpha - \frac{\rho_*}{8\sqrt{d}}\alpha^j \boldsymbol{e}_j$.

Proof: Writing $w = (w_1; w_2)$ with $w_1, w_2 \in \mathbb{R}^d$ for any $w \in \mathbb{R}^{2d}$, consider $f, g_1, g_2 : \mathbb{R}^{2d} \to \mathbb{R}$ defined as:

$$\begin{aligned} f(w) &= \langle \alpha^j \mathbf{e}_j, w_2 - w_1 \rangle, \\ g_1(w) &= \frac{1}{2}\left(\left(\frac{\rho_*}{16\sqrt{d}} \right)^2 - |w_1|^2 \right), \\ g_2(w) &= \frac{1}{2}\left(\left(\frac{\rho_*}{8\sqrt{d}} \right)^2 - \left| w_2 - \frac{3\rho_*}{8\sqrt{d}}\alpha \right|^2 \right). \end{aligned}$$

Then, f, g_1, g_2 are twice continuously differentiable on $\mathbb{R}^{2d}$ and the optimization problem can be re-written as minimizing $f(w)$ over the set $\{w \in \mathbb{R}^{2d} : g_1(w) \geq 0, g_2(w) \geq 0\}$. The proof now follows by noting that the vector $w^* = (w_1^*; w_2^*) \in \mathbb{R}^{2d}$ and the Lagrange multipliers $\lambda_1^* = \frac{16\sqrt{d}}{\rho_*}$ and $\lambda_2^* = \frac{8\sqrt{d}}{\rho_*}$ are the only ones which satisfy the Karush-Kuhn-Tucker second order necessary and sufficient conditions for a strict local solution to this

problem as stated in Theorem 12.5, page 343 and Theorem 12.6, page 345 in Nocedal and Wright (1999). □

A.2.1 Proof of Lemma 4.1

Without loss of generality, we may assume that $r = 1$. Let R_r be $\frac{1}{\sqrt{d}}$ and pick $\delta \in \left(0, \frac{1}{\sqrt{d}}\right)$, $\rho_* = \frac{1}{\sqrt{d}} - \delta$ and $\rho^* = \frac{2d}{1-\delta\sqrt{d}}$. Consider a matrix $Z = (z_1, \ldots, z_d) \in \mathbb{R}^{d\times d}$ with columns $z_1, \ldots, z_d \in \mathbb{R}^d$ and define the function $\tilde{\xi} : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ as

$$\tilde{\xi}(Z) = \begin{vmatrix} \mathbf{e}_1 & z_2^1 - z_1^1 & \cdots & z_d^1 - z_1^1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{e}_d & z_2^d - z_1^d & \cdots & z_d^d - z_1^d \end{vmatrix}$$

where the bars denote the determinant and the equation is written symbolically to express that $\tilde{\xi}(Z)$ is a linear combination of the vectors $\{\mathbf{e}_j\}_{1\le j\le d}$ with the cofactor corresponding to the $(j,1)$-th position as the coefficient of $\mathbf{e}_j$. This is a common notation for "generalized vector products"; see, for instance, Courant and John (1999), Section 2.4.b, page 187 for more details. Since the determinant and all cofactors can be seen as a continuous function on $\mathbb{R}^{d\times d}$, it follows that $\tilde{\xi}$ is continuous on $\mathbb{R}^{d\times d}$. Now choose $\alpha \in \{-1, 1\}^d$ and observe that

$$\begin{aligned} \tilde{\xi}(\alpha^1\mathbf{e}_1, \ldots, \alpha^d\mathbf{e}_d) &= \left(\prod_{j=1}^d \alpha^j\right)\alpha, \\ \left|\tilde{\xi}(\alpha^1\mathbf{e}_1, \ldots, \alpha^d\mathbf{e}_d)\right| &= \sqrt{d}, \\ \langle\tilde{\xi}(\alpha^1\mathbf{e}_1, \ldots, \alpha^d\mathbf{e}_d), \alpha^j\mathbf{e}_j\rangle &= \prod_{k=1}^d \alpha^k \ \ \forall\ j = 1, \ldots, d. \end{aligned}$$

Since $\mathbb{R}^{d\times d}$ has the product topology of the d-fold topological product of $\mathbb{R}^d$ with itself, the continuity of $\tilde{\xi}$ and of $\langle\cdot,\cdot\rangle$ imply that we can find $\rho_\alpha \in \left(0, \frac{1}{\sqrt{d}} - \delta\right)$ such that if $x_j \in$

$B(\alpha^j \mathbf{e}_j, \rho_\alpha)$ for any $j = 1, \ldots, d$, $\beta = \{x_1, \ldots, x_d\}$ and $X_\beta = (x_1, \ldots, x_d)$, then

$$\begin{aligned} \left| |\tilde{\xi}(X_\beta)| - \sqrt{d} \right| &< \delta, \\ \left| \frac{\tilde{\xi}(X_\beta)}{|\tilde{\xi}(X_\beta)|} - \frac{\prod_{1 \le j \le d} \alpha^j}{\sqrt{d}} \alpha \right| &< \delta, \qquad (43) \\ \left| \left\langle \frac{\tilde{\xi}(X_\beta)}{|\tilde{\xi}(X_\beta)|}, x_j \right\rangle - \frac{\prod_{k=1}^d \alpha^k}{\sqrt{d}} \right| &< \delta \ \ \forall \ j = 1, \ldots, d. \qquad (44) \end{aligned}$$

Taking this into account, define

$$\xi_{\alpha,\beta} = \left(\prod_{j=1}^d \alpha^j \right) \frac{\tilde{\xi}(X_\beta)}{|\tilde{\xi}(X_\beta)|}, \text{ and } b_{\alpha,\beta} = \langle \xi_{\alpha,\beta}, x_1 \rangle.$$

From the definition of the function $\tilde{\xi}$ it is straight forward to see that $\langle \xi_{\alpha,\beta}, x_j - x_1 \rangle = 0$ $\forall\, j \in \{1, \ldots, d\}$, so we in fact have

$$x_1, \ldots, x_d \in \mathcal{H}_{\alpha,\beta} := \{x \in \mathbb{R}^d : \langle \xi_{\alpha,\beta}, x \rangle = b_{\alpha,\beta}\}.$$

Moreover, (43) and (44) imply

$$\begin{aligned} \frac{1}{\sqrt{d}} + \delta > b_{\alpha,\beta} > \frac{1}{\sqrt{d}} - \delta > 0, \\ \min_{1 \le j \le d} \left\{ |\xi^j_{\alpha,\beta}| \right\} > \frac{1}{\sqrt{d}} - \delta > 0. \end{aligned}$$

For simplicity, and without loss of generality (the other cases follow from symmetry), we now assume that $\alpha = \mathbf{e}$, the vector of ones. By solving the corresponding quadratic programming problems, it is not difficult to see that

$$\begin{aligned} \rho_* = \frac{1}{\sqrt{d}} - \delta < b_{\alpha,\beta} &= \inf_{\langle \xi_{\alpha,\beta}, x \rangle \ge b_{\alpha,\beta}} \{|x|\} \\ \rho^* = \frac{2d}{1 - \delta\sqrt{d}} > \frac{b_{\alpha,\beta}}{\min_{1 \le j \le d} \{|\xi^j_{\alpha,\beta}|\}} &= \sup_{\substack{\langle \xi_{\alpha,\beta}, x \rangle \le b_{\alpha,\beta} \\ x \ge 0}} \{|x|\}. \end{aligned}$$

For the first inequality see, for instance, Exercise 16.2, page 484 of Nocedal and Wright (1999). For the second one, one must notice that $2\sqrt{d} > \frac{1}{\sqrt{d}} + \delta > b_{\alpha,\beta}$ and that the

optimal value of the optimization problem must be attained at one of the vertices of the polytope $\{x \in \mathbb{R}^d_+ : \langle \xi_{\alpha,\beta}, x\rangle \leq b_{\alpha,\beta}\}$. The latter statement can be derived from the Karush-Kuhn-Tucker conditions of the problem.

The inequalities in the last display imply that $B(0,\rho_*) \subset \mathscr{H}^-_{\alpha,\beta}$ and $\{x \in \mathbb{R}^d : |x| \geq \rho^*\} \cap \mathscr{R}_\alpha \subset \mathscr{H}^+_{\alpha,\beta}$.

Finally, for $x \in B(-\alpha^j \mathbf{e}_j, \frac{1}{2}\rho_\alpha)$ we have $|x + x_j| < \rho_\alpha$ and therefore $\langle \xi_{\alpha,\beta}, x\rangle < -\langle \xi_{\alpha,\beta}, x_j\rangle + \rho_\alpha < \delta - \frac{1}{\sqrt{d}} + \rho_\alpha < 0$. We can then take any $\rho \leq \frac{1}{2}\min_{\alpha\in\{-1,1\}^d}\{\rho_\alpha\}$ to make (i)-(vi) be true. We'll now argue that by making ρ smaller, if required, (vii) also holds.

Let $B_1 = B\left(0, \frac{\rho_*}{16\sqrt{d}}\right)$, $B_2 = B\left(\frac{3\rho_*}{8\sqrt{d}}\alpha, \frac{\rho_*}{8\sqrt{d}}\right)$ and consider the functions $\varphi, \psi : \mathbb{R}^{d\times d} \to \mathbb{R}$ given by

$$\begin{aligned}
\varphi(X) &= \inf_{w_1\in B_1, w_2 \in B_2}\left\{\min_{1\leq j\leq d}\left\{(X(w_2-w_1))^j\right\}\right\},\\
\psi(X) &= \sup_{w_1\in B_1}\left\{\max_{1\leq j\leq d}\left\{(Xw_1)^j\right\}\right\}.
\end{aligned}$$

Both of these functions are Lipschitz continuous with the metric induced by the $\|\cdot\|_2$-norm on $\mathbb{R}^{d\times d}$ with Lipschitz constants smaller than ρ_*. To see this, observe that

$$|X(w_2-w_1) - Y(w_2-w_1)| \leq \|X-Y\|_2 |w_2-w_1| \leq \frac{9}{16}\rho_* \|X-Y\|_2$$

for all $w_1 \in B_1$, $w_2 \in B_2$ and $X, Y \in \mathbb{R}^{d\times d}$. Also, simple algebra shows that $\left|\min_{1\leq j\leq d}\{x^j\} - \min_{1\leq j\leq d}\{y^j\}\right| \leq |x-y|\ \forall\ x, y \in \mathbb{R}^d$. From these assertions, one immediately gets the Lipschitz continuity of φ. Similar arguments show the same for ψ.

Let $\mathscr{I}_\alpha \in \mathbb{R}^{d\times d}$ be the diagonal matrix whose j'th diagonal element is precisely α^j. From Lemma A.3 it is seen that $\varphi(\mathscr{I}_\alpha) = \frac{3\rho_*}{16\sqrt{d}}$. On the other hand, it is immediately obvious that $\psi(\mathscr{I}_\alpha) = \frac{\rho_*}{16\sqrt{d}}$. Using one more time the continuity of ψ and φ and that the topology in $\mathbb{R}^{d\times d}$ is the same as the topology of the d-fold topological product of $\mathbb{R}^d$, for each $\alpha \in \{-1,1\}^d$ we can find r_α for which $X_\beta = (x_1,\ldots,x_d) \in \mathbb{R}^{d\times d}$ and $|x_j - \alpha^j \mathbf{e}_j| < r_\alpha$ for all $j = 1,\ldots,d$ imply $|\psi(X_\beta^{-1}) - \frac{\rho_*}{16\sqrt{d}}| < \frac{\rho_*}{32\sqrt{d}}$ and $|\varphi(X_\beta^{-1}) - \frac{3\rho_*}{16\sqrt{d}}| < \frac{\rho_*}{16\sqrt{d}}$. It follows

that

$$\begin{aligned}
&\inf_{\substack{t\geq 1\\ w_1\in B_1, w_2\in B_2}} \left\{ \min_{1\leq j\leq d} \left\{ \left(X_\beta^{-1}(w_1 + t(w_2 - w_1)) \right)^j \right\} \right\} \\
\geq\ &\inf_{\substack{t\geq 1\\ w_1\in B_1, w_2\in B_2}} \left\{ \min_{1\leq j\leq d} \left\{ \left(tX_\beta^{-1}(w_2 - w_1) \right)^j \right\} \right\} - \sup_{w_1\in B_1} \left\{ \max_{1\leq j\leq d} \left\{ \left(X_\beta^{-1} w_1 \right)^j \right\} \right\} \\
\geq\ &\varphi(X_\beta^{-1}) - \psi(X_\beta^{-1}) \ > \ \frac{\rho_*}{8\sqrt{d}} - \frac{3\rho_*}{32\sqrt{d}} \ = \ \frac{\rho_*}{32\sqrt{d}} > 0.
\end{aligned}$$

The proof is then finished by taking $\rho \leq \min_{\alpha\in\{-1,1\}^d} \{r_\alpha \wedge \frac{\rho_\alpha}{2}\}$. □

A.2.2 Proof of Lemma 4.2

Assume again, without loss of generality, that $r = 1$. Lemma 4.1 (ii) and (vi) imply that $x_{\alpha^j j}, x_{-\alpha^j j} \in \{x \in \mathbb{R}^d : \langle x, \xi_\alpha\rangle \leq b_\alpha\}$ for any $j = 1,\ldots,n$ and any $\alpha \in \{-1,1\}^d$. It follows that, in addition to being convex, $\cap_{\alpha\in\{-1,1\}^d}\{x \in \mathbb{R}^d : \langle \xi_\alpha, x\rangle \leq b_\alpha\}$ contains $\{x_{\pm 1},\ldots,x_{\pm d}\}$ and hence it must contain K. For the other contention, take $x \in \cap_{\alpha\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle \leq b_\alpha\}$ with $x \neq 0$ and any $\alpha \in \{-1,1\}^d$ for which $x \in \mathscr{R}_\alpha$. Then, $\langle \xi_\alpha, x\rangle > 0$ for otherwise we would have

$$\kappa x \in \mathscr{R}_\alpha \setminus \mathscr{H}_\alpha^+ \ \ \forall\ \kappa \geq 0$$

which is impossible by (v) in Lemma 4.1. Thus, $\mathscr{J}_x = \{\alpha \in \{-1,1\}^d : \langle \xi_\alpha, x\rangle > 0\} \neq \emptyset$ and we can define

$$r_x = \min_{\alpha\in\mathscr{J}_x} \left\{ \frac{b_\alpha}{\langle \xi_\alpha, x\rangle} \right\} \ \text{ and } \alpha_x = \operatorname*{argmin}_{\alpha\in\mathscr{J}_x} \left\{ \frac{b_\alpha}{\langle \xi_\alpha, x\rangle} \right\}.$$

Note that $r_x \geq 1$. Since β_{α_x} is a basis, there is $\theta \in \mathbb{R}^d$ such that $r_x x = \theta^1 x_{\alpha_x^1 1} + \ldots + \theta^d x_{\alpha_x^d d}$. But then,

$$b_{\alpha_x} = \langle r_x x, \xi_{\alpha_x}\rangle = \sum_{k=1}^d \theta^k \langle x_{\alpha_x^k k}, \xi_{\alpha_x}\rangle = b_{\alpha_x} \sum_{k=1}^d \theta^k$$

where the last equality follows from (ii) of Lemma 4.1 and therefore $\theta^1 + \ldots + \theta^d = 1$. Now assume that $\theta^j < 0$ for some $j \in \{1,\ldots,d\}$ and set $\gamma_x \in \{-1,1\}^d$ with $\gamma_x^k = \alpha_x^k$ for $k \neq j$ and

$\gamma_x^j = -\alpha_x^j$. But then, $\sum_{k\neq j}\theta^k = 1-\theta^j > 1$, $\langle x_{\alpha_x^k k}, \xi_{\gamma_x}\rangle = b_{\gamma_x}$ for $k \neq j$ and $\langle x_{\alpha_x^j j}, \xi_{\gamma_x}\rangle < 0$ by (ii) and (vi) in Lemma 4.1. Therefore,

$$\begin{aligned}\langle r_x x, \xi_{\gamma_x}\rangle &= \theta^j \langle x_{-\alpha_x^j j}, \xi_{\gamma_x}\rangle + \sum_{k\neq j}\theta^k \langle x_{\alpha_x^k k}, \xi_{\gamma_x}\rangle &(45)\\ &> \sum_{k\neq j}\theta^k \langle x_{\alpha_x^k k}, \xi_{\gamma_x}\rangle > b_{\gamma_x} &(46)\end{aligned}$$

which is impossible because it contradicts the definition of r_x. Hence, $\theta \geq 0$ and we have $r_x x \in Conv\left(\beta_{\alpha_x}\right)$. Note that since 0 belongs in the interior of $\cap_{\alpha\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle \leq b_\alpha\}$, there there is $\kappa > 0$ such that $-\kappa x \in \cap_{\alpha\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle \leq b_\alpha\}$. Applying the same arguments as before to $-\kappa x$ instead of x, we can find $\tilde{r}_x > 0$ and $\tilde{\alpha}_x \in \{-1,1\}^d$ such that $-\tilde{r}_x x \in Conv\left(\beta_{\tilde{\alpha}_x}\right)$. It follows that $-\tilde{r}_x x, r_x x \in K$ and therefore $0, x \in K$ since $r_x \geq 1$. Hence, we have proved (i).

To prove (ii), note that $A := \cap_{\alpha\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle < b_\alpha\}$ is open and, by (i), it is contained in K. Thus, $A \subset K^\circ$. That $K^\circ \subset A$ follows from the fact that if $x \in K \setminus A$, then $\langle \xi_\alpha, x\rangle = b_\alpha$ for some $\alpha \in \{-1,1\}^d$, which implies that $B(x,\tau) \cap \mathrm{Ext}(K) \neq \emptyset$ for all $\tau > 0$ and hence $x \notin K^\circ$.

It is then obvious that (iv) follows from the identity $\partial K = \overline{K} \setminus K^\circ$ and the fact that K is closed.

Pick any $\alpha \in \{-1,1\}^d$ and observe that (ii) and (vi) from Lemma 4.1 imply that for any $\gamma \in \{-1,1\}^d$ we have

$$\langle \xi_\gamma, x_{\alpha^k k}\rangle \begin{cases} = b_\gamma & \text{if } \gamma^k = \alpha^k \\ < 0 \leq b_\gamma & \text{if } \gamma^k = -\alpha^k \end{cases}$$

which by (iv) of this lemma show that

$$x_{\alpha^j j} \in \{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle = b_\alpha\} \cap \left(\cap_{\gamma\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\gamma, w\rangle \leq b_\gamma\}\right)$$

for all $\alpha \in \{-1,1\}^d$ and $j = 1,\ldots,d$. Since the sets on the right-hand side of the last display are all convex we can conclude that

$$Conv\left(x_{\alpha^1 1},\ldots,x_{\alpha^j j}\right) \subset \{w \in \mathbb{R}^d : \langle \xi_\alpha, w\rangle = b_\alpha\} \cap \left(\cap_{\gamma\in\{-1,1\}^d}\{w \in \mathbb{R}^d : \langle \xi_\gamma, w\rangle \leq b_\gamma\}\right)$$

for all $\alpha \in \{-1,1\}^d$. Thus, $\bigcup_{\alpha\in\{-1,1\}^d} Conv\left(x_{\alpha^1 1},\ldots,x_{\alpha^j j}\right) \subset \partial K$. Finally, take $x \in \partial K$. Then, there is $\alpha_x \in \{-1,1\}^d$ such that $\langle \xi_{\alpha_x}, x\rangle = b_{\alpha_x}$. Since β_{α_x} is a basis we can again find $\theta \in \mathbb{R}^d$ such that $x = \theta^1 x_{\alpha_x^1 1} + \ldots + \theta^d x_{\alpha_x^d d}$. Just as before, $\langle \xi_{\alpha_x}, x_{\alpha_x^j j}\rangle = b_{\alpha_x}$ implies that $\sum \theta^j = 1$. And again, if $\theta^j < 0$ for some j, we can take $\gamma_x \in \{-1,1\}^d$ with $\gamma_x^k = \alpha_x^k$ for $k \neq j$ and $\gamma_x^j = -\alpha_x^j$ and arrive at a contradiction with similar arguments to those used in (45) and (46). This shows that $x \in Conv\left(\beta_{\alpha_x}\right)$ and completes the proof as (v) and (vi) are direct consequences of $(i)-(iv)$ and Lemma 4.1. □

A.2.3 Proof of Lemma 4.3

Let $r \in (0, \frac{1}{d-2})$ if $d \geq 3$ and $r > 0$ if $d \leq 2$. Since the geometric properties of any rectangle depend only on the direction and magnitude of the diagonal, we may assume without loss of generality that $b > 0$ and that $a = \frac{r}{1+r} b$. This is because we can define $\tilde{b} = (1+r)(b-a) > 0$ and $\tilde{a} = a - r(b-a)$ to obtain $[a,b] = \tilde{a} + \left[\frac{r}{r+1}\tilde{b}, \tilde{b}\right]$. For any $\alpha \in \{-1,1\}^d$, define $\alpha_j = \alpha - 2\alpha^j \mathbf{e}_j \in \mathbb{R}^d$ and $w_\alpha = z_\alpha + r(z_\alpha - z_{-\alpha})$. Additionally, define the functions $\psi_\alpha, \varphi_\alpha : \mathbb{R}^{d\times d} \times \mathbb{R}^d \to \mathbb{R}$ by

$$\begin{aligned} \psi_\alpha(\Theta,\theta) &= \langle \mathbf{e}, \Theta(z_\alpha - \theta)\rangle \\ \varphi_\alpha(\Theta,\theta) &= \min_{1\leq j\leq d} \left\{ (\Theta(z_\alpha - \theta))^j \right\}. \end{aligned}$$

Considering $\mathbb{R}^{d\times d}$ with the topology generated be the $\|\cdot\|_2$ norm and $\mathbb{R}^{d\times d} \times \mathbb{R}^d$ with the product topology, it is easily seen that both functions defined in the last display are continuous. Now, let $W_\alpha \in \mathbb{R}^{d\times d}$ be the matrix whose j'th column is precisely $w_{\alpha_j} - w_\alpha$. It is not difficult to see that $\psi_\alpha(W_\alpha^{-1}, w_\alpha) = \frac{dr}{1+2r} < 1$ and $\varphi_\alpha(W_\alpha^{-1}, w_\alpha) = \frac{r}{1+2r} > 0$. For instance, one can check that for $\alpha = -\mathbf{e}$, one has $w_\alpha = 0$ and $w_{\alpha_j} = \frac{1+2r}{1+r} b^j \mathbf{e}_j$ and the result is now evident. By symmetry, the same is true for any $\alpha \in \{-1,1\}^d$. Therefore, for any $\alpha \in \{-1,1\}^d$ there is ρ_α such that whenever $|x_{\alpha_j} - w_{\alpha_j}| < \rho_\alpha \ \forall\ j = 1,\ldots,d$ and X_α is

the matrix whose j'th column is $x_{\alpha_j} - x_\alpha$, we get

$$\psi_\alpha(X_\alpha^{-1}, x_\alpha) < 1, \tag{47}$$

$$\varphi_\alpha(X_\alpha^{-1}, x_\alpha) > 0. \tag{48}$$

Letting $\rho = \min_{\alpha\in\{-1,1\}^d}\{\rho_\alpha\}$ completes the proof as (47) and (48) imply $z_\alpha \in Conv\left(x_\alpha, x_{\alpha_1}, \ldots, x_{\alpha_d}\right)^\circ$. □

References

Allon, G., Beenstock, M., Hackman, S., Passy, U., and Shapiro, A. (2007). Nonparametric estimation of concave production technologies by entropic methods. *J. Appl. Econometrics*, 4:795–816.

Banker, R. and Maindiratta, A. (1992). Maximum likelihood estimation of monotone and concave production frontiers. *J. Productiv. Anal.*, 3:401–415.

Beresteanu, A. (2007). Nonparametric estimation of regression functions under restrictions on partial derivatives. Available at: http://econ.duke.edu/ arie/shape.pdf (unpublished manuscript).

Birke, M. and Dette, H. (2006). Estimating a convex function in nonparametric regression. *Scand. J. Statist.*, 34:384–404.

Boland, N. L. (1997). A dual-active-set algorithm for positive semi-definite quadratic programming. *Math. Program.*, 78:1–27.

Bronšteĭn, E. M. (1978). Extremal convex functions. *Sibirsk. Mat. Zh.*, 19:10–18.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616.

Brunk, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference*, pages 177–197. Camridge University Press, New York, NY, USA.

Chung, K. L. (2001). *A Course in Probability Theory*. Academic Press, San Diego, CA, USA.

Conway, J. (1985). *A Course in Functional Analysis*. Springer-Verlag, New York, NY, USA.

Courant, R. and John, F. (1999). *Introduction to Calculus and Analysis, Vol. II/1*. Springer, New York, NY, USA.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.*, 4:254–270.

Cule, M., Samworth, R., and Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B (to appear)*.

Dudley, R. M. (1977). On second derivatives of convex functions. *Math. Scand.*, 41:159–174.

Folland, G. (1999). *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, NY, USA.

Grenander, U. (1956). On the theory of mortality measurement. *Skan. Aktuarietidskr, Part II*, 39:125–153.

Groeneboom, P., Jongbloed, G., and Wellner, J. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29:1653–1698.

Hanson, D. L. and Pledger, G. (1976). Consistency in concave regression. *Ann. Statist.*, 4:1038–1050.

Harville, D. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, New York, NY, USA.

Hildreth, C. (1954). Estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.*, 49:598–619.

Johansen, S. (1974). The extremal convex functions. *Math. Scand.*, 41:61–68.

Kapoor, S. and Vaidya, P. M. (1986). Fast algorithms for convex quadratic programming and multicommodity flows. *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 147–159.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econom. J.*, 11:308–325.

Luenberger, D. (1984). *Linear and Nonlinear Programming*. Addison-Wesley Publishing Company, Reading, MA, USA.

Maindiratta, A. and Sarath, B. (1997). On the consistency of maximum likelihood estimation of monotone and concave production frontiers. *J. Productiv. Anal.*, 8:239–246.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.*, 19:741–759.

Matzkin, R. L. (1991). Semiparametric estimation of monotone concave utility functions for polychotomous choice models. *Econometrica*, 59:1351–1327.

Matzkin, R. L. (1993). Nonparametric identification and estimation of polychotomous choice models. *J. Econometrics*, 58:137–168.

Mehrotra, S. and Sun, J. (1990). An algorithm for convex quadratic programming that requires $o(n^{3.5}l)$ arithmetic operations. *Math. Oper. Res.*, 15:342–363.

Moreau, J. (1962). Decomposition orthogonal d'un espace hilbertien selon deux cones mutuellement polaires. *C. R. Acad. Sci. Paris*, 255:238–240.

Nocedal, J. and Wright, S. (1999). *Numerical Optimization*. Springer, New York, NY, USA.

Rockafellar, T. R. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ, USA.

Schuhmacher, D. and Dümbgen, L. (2010). Consistency of multivariate log-concave density estimators. *Statist. Probab. Lett.*, 80:376–380.

Schuhmacher, D., Hüsler, A., and Dümbgen, L. (2009). Multivariate log-concave distributions as a nearly parametric model. Tech. rep., University of Bern. Available at: http://arxiv.org/abs/0907.0250.

Seregin, A. and Wellner, J. (2009). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist. (to appear)*. Available at: http://arxiv.org/abs/arXiv:0911.4151v1.

Song, W. and Zhengjun, C. (2004). The generalized decomposition theorem in banach spaces and its applications. *J. Approx. Theory*, 129:167–181.

Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, USA.

Varian, H. (1982a). The nonparametric approach to demand analysis. *Econometrica*, 50:945–973.

Varian, H. (1982b). The nonparametric approach to production analysis. *Econometrica*, 52:579–597.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge, UK.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.*, 30:528–555.